Load Profiling on Time and Spectral Domain
From Big Data to Smart Data

Li, Ran

Award date:
2015

Awarding institution:
University of Bath

Load Profiling on Time and Spectral Domain: From Big Data to Smart Data

By

Ran Li

BEng, SMIEEE

The thesis submitted for the degree of

Doctor of Philosophy

in

The Department of
Electronic and Electrical Engineering
University of Bath

December 2014

Signature:…… Date:……

Contents

Contents ……...ii
Abstract ...……...vi
Acknowledgements viii
List of Figures ix
List of Tables.....xi
List of Abbreviations.....xii
Chapter 1 Introduction.....1
1.1 Overview.....2
1.1.1 Drivers: Climate Change and Low Carbon Techniques 2
1.1.2 Practical Constraints: Limited Visibility 3
1.2 Load Profiling 4
1.2.1 Load Profiling in Small Data Era.....4
1.2.2 Load Profiling in Big Data Era 5
1.3 Research Challenges and Objectives 7
1.3.1 Load Profiling for LV Networks.....7
1.3.2 Load Profiling for Individual Customers.....8
1.4 Research Contributions.....9
1.5 Thesis Layout.....13
Chapter 2 Review of Load Profiles 15
2.1 Introduction.....16
2.2 Development of Load Profiles 16
2.2.1 Definition of Load Profiles 16
2.2.2 Factors of load profiles variation 16
2.2.3 Applications of Load Profiles 17
2.2.4 Load Profiles across the World.....18
2.3 Load Profiling Methods 21
2.3.1 Customer Classification and Load Shape Clustering.....21
2.3.2 Load Magnitude Scaling 24

2.4 Chapter Summary 26
Chapter 3 Evaluation of Engneering Load Profiles in the UK 27
3.1 Introduction 28
3.2 Power synthesis method for Load Profile Test 28
3.3 Customer Classification 29
3.3.1 Customer size 29
3.3.2 Classify Customers into Eight Classes 31
3.4 Estimation of Substation Profiles 32
3.5 Derivation of Metered Substation Profiles and Comparisons 33
3.5.1 Substation power profiles 33
3.5.2 Comparisons 34
3.5.3 Other Potential Causes 35
3.5.4 Test on Low Voltage Level and Individual Customers 36
3.6 Chapter Summary 38
Chapter 4 Time-series Load Profiling for LV Networks: Clustering and Classification 39
4.1 Introduction 40
4.2 Problem and Proposed Solution Statement 40
4.3 LV Network Templates Project 43
4.4 Overall Flowchart of the Methodology 45
4.5 Clustering and Classification 46
4.5.1 Hierarchical Clustering for Load Profiles 46
4.5.2 K-Means Clustering for Cluster Number Determination 49
4.5.3 Multinomial Logistic Regression for Classification 50
4.6 Implementation 51
4.6.1 Normalization 51
4.6.2 Determining Cluster Number 53
4.6.3 Classification Tool 54
4.7 Demonstration and Results 55
4.7.1 Number of clusters 55
4.7.2 Clusters and Normalised Templates 56
4.7.3 Classification Tool 61
4.8 Chapter Summary 61

Chapter 5 Time-series Load Profiling for LV Networks: Peak Load Estimation 62
5.1 Introduction 63
5.2 Problem and Proposed Solution Statement 63
5.3 Rationale of Contribution Factor 65
5.3.1 Latency in LV Substation Class 65
5.3.2 Contribution Factor 66
5.3.3 Illustration of Contribution Factor 66
5.4 CWCR Based Scaling Method 67
5.5 Mathematical Formulation 68
5.5.1 Clusterwise Weighted Constrained Regression 69
5.5.2 Cross Validation 70
5.5.3 Tests of P-Q Method and C-D Method 71
5.6 Results 72
5.6.1 Fitness Comparison 72
5.6.2 Cross Validation 76
5.7 Discussion on the Use of Network Templates 78
5.8 Chapter Summary 80
Chapter 6 Spectral Load Profiling for Individual Customer: Feature Extraction 81
6.1 Introduction 82
6.2 Problem and Proposed Solution Statement 82
6.3 Spectral Analysis and Data Description 85
6.4 Decomposition and Reconstruction 86
6.4.1 Discrete Fourier Transform 87
6.4.2 Discrete Wavelet Transform 88
6.5 Assessment Method 90
6.6 Results for Smart Metering Data 91
6.6.1 Individual Customer 92
6.6.2 Load Characterisation 92
6.6.3 Data Compression 94
6.7 Assessment over Different Aggregation Levels 97
6.7.1 Monthly Averaged Load Profiles 97
6.7.2 LV Substation 99

6.8 Chapter Summary 100
Chapter 7 Spectral Load Profiling for Individual Customer: Multi-resolution Clustering 101
7.1 Introduction 102
7.2 Problem and Proposed Solution Statement 103
7.3 Clustering Techniques 105
7.3.1 GMM 105
7.3.2 Parameter estimation by EM algorithm 106
7.3.3 X-means 107
7.4 Multi-resolution Clustering 108
7.4.1 Time-series clustering 108
7.4.2 MRC 110
7.5 Classifications 114
7.6 Results 116
7.6.1 TCs 116
7.6.2 Comparisons 119
7.7 Chapter Summary 122
Chapter 8 Conclusion 123
Chapter 9 Future Work 130
Appendix. A 134
Appendix. B 139
Appendix. C 141
Appendix. D 151
Appendix. E 154
Publications 159
Reference 194

Abstract

With the promotion of demand side responses (DSRs) and low carbon technologies (LCTs), there is a growing interest in visualising the demand information at individual consumer and low voltage (LV) network level, where demands are less aggregated and highly volatile. Yet, traditional load profiling techniques, which are carried out on small data, are struggling to meet the requirements on accuracy and granularity. This thesis contributes to this area by extending traditional load profiling to a big-data context, where refined load profiles (smart data) can be extracted by two novel load profiling techniques for LV networks and individual consumers. The refined load profiles aim to: i) economically visualise LV networks with limited smart-grid monitoring data; ii) transform the smart metering data into a high-detail granular representation of the customers' daily demand.

For the LV networks, this thesis develops a novel concept, LV network templates, which aim to visualise the LV networks in a cost-effective manner. A novel three-stage load profiling method is proposed as: clustering, classification and scaling. By using statistical time-series analysis, three steps are undertaken: i) cluster a vast amount of load data according to their load shapes; ii) classify un-monitored substations to the most similar cluster without sample metering; iii) and also scale them to the right magnitude without sample metering. Through this method, limited representative monitoring data can be used to develop a library of typical load profiles for un-monitored networks, thus saving the cost of extensive monitoring for every single substation. In addition, it is the first load profiling method that can accurately express both load shapes and magnitudes for LV networks.

Regarding the customer's demand representation, the developed time-series analysis needs to be updated due to the volatile and uncertain nature of smart metering data, including inter-related factors such as overall load shapes, sudden spikes and magnitudes. Therefore, an innovative spectral load profiling is proposed to decompose these factors into different spectral levels, characterised by spectral features. By analysing the extracted features on each spectral level separately through

multi-resolution analysis, the interference among different factors can effectively be prevented. The proposed method, for the first time, is able to fully capture the energy characteristics at the household level.

The developed LV network load templates provide an economical but straightforward way to quantify the available headroom of unmonitored substations over time, providing quantitative information for distribution network operators to integrate LTCs at the minimal costs. The spectral load profiling gives an insight into customer's energy behaviours with high granularity and accuracy. It can support the customer-specified DSR, tariff design, smart metering validation and load forecasting.

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor Prof. Furong Li for her invaluable guidance on my research and consistent support throughout the course of my study.

I would like to express my thanks to Western Power Distribution and University of Bath, who jointly funded my research. I would like to express my gratitude to everyone involved in the Low Voltage Network Template Project. Particularly, I would like to thank Dr. Gavin Shaddick and Dr. Haojie Yan for tremendous guidance on my statistics and programming. I would also like to thank Dr. Nathan Smith for his generous advices on signal processing.

I would like to express my heartfelt gratefulness for my previous fellow colleagues, Dr. Bo Li, Dr. Yan Zhang, Dr. Chenchen Yuan and Dr. Zhimin Wang for constructive suggestions and advices. I am also genuinely thankful to my current colleagues, Mr. Fan Yi, Mr. Jiangtao Li and Miss. Lin Zhou for sharing knowledge and resources. In particular, I would like to thank Dr. Chenghong Gu, Dr. Ignacio Hernando Gil, Mr. Zhipeng Zhang and Miss. Chen Zhao for their assistances on this thesis.

And, I would also thank all my friends in the University of Bath, including Dr.Shuang Yu, Dr. Chao Gao, Mr. Hualei Wang, Mr. Maomao Zhang and Lido Chef for their support over years.

Last but not least, I would like to take this opportunity to express my ultimate gratitude to my family. I am sincerely thankful to my parents, Qiying and Xiuti, who are my lifelong role models. I am grateful to my beloved wife, Dujuan, who gives me endless support and encouragement.

List of Figures

Figure 1-1 Sketch showing how anticipated modelling error varying with i) feature extraction; ii) classification 11
Figure 1-2 Overview of the research 12
Figure 2-1 Domestic profile classes for winter Wednesday 20
Figure 2-2 Non-Domestic profile classes for winter Wednesday 20
Figure 2-3 Non-Domestic Maximum Demand profile classes for winter Wednesday 20
Figure 2-4 Flow charts of classification methods by fixed data 22
Figure 2-5 Flow charts of classification methods by fixed data 23
Figure 3-1 Overall process of power synthesis method 29
Figure 3-2 Estimated substation profile for winter Wednesday 32
Figure 3-3 System map of dowlishford substation 33
Figure 3-4 Comparison between metered and estimated data for winter Wednesdays 34
Figure 3-5 Comparison between metered and estimated data for summer Wednesdays 35
Figure 3-6 Comparison between similar LV substations on 11th July 37
Figure 3-7 Comparison between similar LV substations on 11th July 38
Figure 4-1 Geographical areas of LV Network Template project 44
Figure 4-2 Flow chart of the methodology 45
Figure 4-3 Example of a dendogram 49
Figure 4-4 Annual load of a selected substation by normalised I method 52
Figure 4-5 Within-group sum of squares errors against number of clusters 54
Figure 4-6 LV substation template and within-cluster variability 58
Figure 4-7 Templates and standard deviations for clusters 1 and 2 58
Figure 4-8 Templates and standard deviations for clusters 3 and 4 59
Figure 4-9 Templates and standard deviations for clusters 5 and 6 59
Figure 4-10 Templates and standard deviations for clusters 7 and 8 60
Figure 4-11 Templates and standard deviations for clusters 9 and 10 60
Figure 5-1. Load profiles of a typical domestic customer, cluster 5 and cluster 6 65
Figure 5-2. Individual customer contribution to substation peak 67
Figure 5-3 Ratio of metered and estimated peak by industry P-Q method 72
Figure 5-4 Metered and estimated peaks by C-D method (coefficients from OLS regression) 72
Figure 5-5 Metered and estimated peaks by CWCR 73
Figure 5-6 Error comparison of different methods 73
Figure 5-7 Ratio of metered and estimated peaks by P-Q method 76
Figure 5-8 Ratio of metered and estimated peaks by cluster regression (cluster 4) 77
Figure 5-9 Load profile template of cluster 1 79
Figure 6-1 Comparison between traditional TLPs and smart metered load profiles (Data from Irish Smart Metering Project) 83
Figure 6-2 Load profile decomposition by DFT 88

Figure 6-3 Multi-resolution analysis by DWT.....89
Figure 6-4 Load profile decomposition by DWT 89
Figure 6-5 Daily load profiles of customer 1002 in July 2012.....92
Figure 6-6 Decomposition components from DWT customer 1002 93
Figure 6-7 Periodical sinusoidal components from DFT.....94
Figure 6-8 Accumulated energy by keeping different number of coefficients 95
Figure 6-9 Percentage of customers who can be reconstructed under the threshold error with different data size 96
Figure 6-10 Daily individual load profile and reconstructions by reduced DFT and DWT coefficients 98
Figure 6-11 Monthly average load profile and reconstructions by reduced DFT and DWT coefficients 98
Figure 6-12 Percentage of customers who can be reconstructed under the threshold error with different data size (monthly) 98
Figure 6-13 Average minimum data required to reconstruct load profiles from DFT and DWT coefficients for different customer groups (PMEI, MME, MAPE< 5% and PTE<2 hours).....99
Figure 7-1 Conventional load profile clustering process.....108
Figure 7-2 Problems with time-series clustering: magnitude difference within clusters 109
Figure 7-3 Problems with time-series clustering: time difference in spikes 109
Figure 7-4 Problems with time-series clustering: uncertainties between days ..110
Figure 7-5 Overall methodology of multi-resolution clustering 111
Figure 7-6 Two-stage GMM and X-means clustering implemented in MRC 113
Figure 7-7 TC and members of cluster 1 on A level 117
Figure 7-8 TC and members of cluster 2 on A level 117
Figure 7-9 TC and members of cluster 1 on D2 level 117
Figure 7-10 TC and members of cluster 2 on D2 level 118
Figure 7-11 TC and members of cluster 1 on D1 level 118
Figure 7-12 TC and members of cluster 2 on D1 level 119
Figure 7-13 Posterior probability spectrum (A level) of customer 1609 over a month 119
Figure 7-14 Comparison between smart metering data of customer 1609 on 26/08/2009 (black) and three load profiling methods: UK TLP (green), K-means(blue), MRC (red).....121
Figure 7-15 Uncertainties between days are resolved by MRC onAlevel.....122
Figure 7-16 Time delay of spikes are resolved by MRC on A level.....122

List of Tables

Table 2-1 Load Classes and Their Descriptions.....19
Table 3-1 Number of households served by dowlishford substation in 2011 30
Table 3-2 Households numbers of eight typical load profiles.....32
Table 4-1 Predictive accuracy of classification with various cluster number 56
Table 4-2 Ten LV network templates.....57
Table 5-1 Error Analysis of different estimation methods (kW).....74
Table 5-2 Goodness-to-fit Comparisons.....75
Table 5-3 R Squared Error for all clusters and seasons.....76
Table 5-4 Comparison of Cross Validation on P-Q and CWCR.....77
Table 5-5 Low carbon capacity for LV substations in cluster 1.....79
Table 5-6 Load factors for all clusters and seasons 80
Table 6-1 DFT coefficients of a sampled load profile.....90
Table 7-1 Number of clusters of each group and decomposition level 116
Table 7-2 Classification of sampled customer by different load profiling methods 120
Table 7-3 Comparison between smart metering load profiles and three load profiling methods (sample size=2994).....121

List of Abbreviations

Bayesian Information Criterion	BIC
Carbon Dioxide	CO_2
Department of Energy & Climate Change	DECC
Distribution Network Operators	DNOs
Demand Side Response	DSR
Discrete Fourier Transforms	DFT
Discrete Wavelet Transforms	DWT
Electric vehicles	EV
Heat Pumps	HP
Great Britain	GB
Gaussian Mixture Model	GMM
High Voltage	HV
Low Voltage	LV
Low Carbon Technologies	LCTs
Multinomial Logistic Regression	MLR
Multi-resolution Clustering	MRC
Industrial and Commercial	I&C
Office of Gas and Electricity Markets	Ofgem
Typical Load Profiles	TLPs
Photovoltaic	PV
United Kingdom	UK
Western Power Distribution	WPD

Chapter 1

Introduction

THIS chapter describes the background, motivations, objectives, challenges and contributions of this work. It also presents a structure of the thesis.

1.1 Overview

1.1.1 Drivers: Climate Change and Low Carbon Techniques

Climate change has recently become one of the most complicated challenges among all the environment related issues. Actually, the Earth's surface has warmed by about 0.8°C since 1900 [1]. The massive amount of carbon dioxide (CO_2) emissions, levels of which in the atmosphere have increased by about 40% since the beginning of the industrial revolution, explain this warming through their enhancement of the natural greenhouse effect [2]. Most countries and governments have made agreements on reduction of the greenhouse gasses and work on achieving their own targets. The UK government is committed to cutting off greenhouse gas emissions by at least 80% by 2050, relative to 1990 levels [1].

Under this long term plan, low carbon technologies (LCTs) such as renewable generation and low-carbon appliances have seen a significant introduction throughout the UK. Renewable generations, such as solar energy, wind energy and bioenergy are encouraged by the policy to provide clean energy for the UK [3, 4]. In the past two years (2012-2013), renewable generation increased by 30% to 53.7 TWh, which contributes up to 14.9% of the total UK electricity generation [5]. There is also a corresponding interest in low-carbon appliances such as electric vehicles (EVs) and heat pumps (HPs) especially at the low voltage (LV, 415V) customer side. The number of EVs in the UK will reach 0.9 million by 2020 based on a conservative estimate [6].

The extensive integration of LCTs will potentially bring unprecedented challenges to the electricity network planning and operation. In the UK alone, the cost of integrating these technologies using the traditional approaches is estimated to be around £200 billion by 2020 [7]. New approaches for network planning and customer operation are therefore required in order to mitigate the network pressure and to further improve the power system efficiency. Two promising alternative solutions are the implementation of low carbon smart network (LCN) [8] and demand side response (DSR) [9]. The concepts are listed as follows.

- The concept of low carbon network describes a system which can efficiently

integrate and manage LTCs meanwhile maintaining reliable and affordable grid service. An efficient LCN will take account the voltage and thermal headrooms of the network, and be able to make the full usage of these for operating and planning the system and the LCTs to minimise the integration costs. The Office of Gas and Electricity Markets (Ofgem) in the UK has established the LCN fund which supports a series of projects [10] to investigate alternative technology, operation and commercial arrangements for improving the efficiency of the distribution networks.

- DSR is a program that encourages LV-end power consumers to vary demand according to price or control signals in order to avoid periods of high energy price and high network demand [8]. It can be realised through the utilisation of either the customers' time-movable electricity consumption or LCTs. DSR can lead to financial benefits by energy cost reduction and deferring network investment. In addition, DSR can increase the efficiency of networks and LCTs, which leads to a potential CO_2 emission reduction.

One of the key factors of realising the LCN and DSR is load visibility, i.e. to identify repeatable patterns of demand and their location in the system. Accurate knowledge of the LV networks' conditions can help DNOs with efficient network operation and planning in the presence of significant LCTs, such as when and where to integrate certain type of LCTs without triggering network reinforcement. Also, the strategies development and incentives design of DSR are also based on understanding customers' electricity energy behaviours.

1.1.2 Practical Constraints: Limited Visibility

However, the main obstacle for LCN and DSR now is the limited visibility of the power flow in LV networks and individual customers' load profiles.

- There are options to visualise the power flows in LV networks, but all of them face practical constraints: i) existing industrial methods follow an indirect bottom-up approach, which aggregate all connected end-users' load to estimate the demand of LV networks. However, these methods cannot meet the requirements for LCN due to their inaccuracy [5]; ii) as the most direct approach, extensive installation of

monitoring devices on LV networks improves the accuracy, but can be prohibitively expensive. In the UK, there are more than 900,000 LV substations [11] and it would cost over £2 billion to install metering devices, not to mention the data acquisition equipment and daily data management.

- Individual customer's load profiles can be extremely volatile and uncertain. The uncertainty is shown by two main aspects: the load profile variance between different customers and the load profile variance of the same customer over different days. The understanding of customer energy usage behaviour is very limited and existing load profiling methods cannot express the load behaviours of mass customers in fine details because they were designed to express aggregated load shapes. Consequently, DSR strategies based on the traditional load profiles may not guide individual customers to effectively support energy and network needs. Worse, it could further aggravate energy or network problems.

1.2 Load Profiling

Load profiles have been traditionally used to increase the visibility of the LV networks and mass customers [12]. Due to the diversity of customer types and extensiveness of distribution networks, it is impractical to collect load information for every customer continuously over time. A typical load profile (TLP), which is a graph showing the load variation versus time [13], has been adopted to represent a group of similar customers.

Load profiling is the process of developing load profiles. The common approach of load profiling is: i) for mass customers, to assemble similar customers into pre-defined group, and to study the typical load variations within each group [14]; ii) for LV networks, it is a common practice to aggregate the customers load up to network level as an approximation [15].

1.2.1 Load Profiling in Small Data Era

In the UK, load profiles were originally created for electricity settlement when the electricity market first opened up in the 1990s [12]. The purpose was to avoid expensive cost of installing smart meters at every individual household. Customers

with maximum power demand below 100 kW were roughly divided into 8 classes and represented by TLPs [14]. The detailed eight profile classes and their characteristics will be introduced in Chapter II.

Although the traditional load profiles have served the industry well for decades, they are unable to support either LCN or DSR in the new environment, which critically requires more accurate and granular load information. Traditional load profiling methods are challenged especially for the following limitations:

i) LV networks: to visualise the LV networks, traditional load profiling methods usually follow the bottom-up approach, which aggregates customers' load profiles up to the substation level. However, these indirect methods do not represent the diversity in customer and network characteristics. They may be easily distorted at any link in the chain: the inaccuracy of customer load profile itself, misclassification of customers, different network structures and variance in loading levels.

ii) Individual customer: traditional load profiles were developed from small sample size, pre-defined classifications and applied to roughly approximate a group of customers over a season. Error arises due to a large variance within each customer class, and also between days. It is unable to express the volatile and uncertain load profiles of individual customers on individual days. Also, customers are classified according to some pre-knowledge, which does not necessarily indicate similar load profiles.

1.2.2 Load Profiling in Big Data Era

Network monitors and smart meters now provide new opportunity to enhance LV network visibility and to understand customer energy usage patterns. Smart meters are the next generation of gas and electricity meters. The Department of Energy & Climate Change (DECC) has aimed to install smart meters for all homes and small businesses by 2020 [16]. The transition will involve rolling out over 53 million smart meters [17].

In spite of the rich benefits from real-time load information, smart meters also bring unprecedented challenges of high cost and big data:

i) In terms of the high cost, to cover nearly 1 million substations and 53 million mass customers in the UK [16], nearly 60 million smart meters or monitoring devices would be required. The smart meters roll-out will cost more than £13 billion exclude the cost of data management.

ii) Based on the IBM big data trial [18] with the top data algorithms and processor in the world, it is estimated that the data of 60 million smart meters can produce 7 TB over only one month, which can be compressed to 2.4 TB for storage. However, to load and decompress the data will take nearly 2 hours. As the data will be used for various precise applications including settlement, customer billing, DSR and tariff design, it would be essential to take up data validation, pre-processing, missing data estimation etc., which will cost extra time and resources.

A blind search in big data might increase the burden of power system and curtail the efficiency instead. It is important to extract the meaningful load information, which is targeted as *smart data* in this thesis. New load profiling methods become essential with two main objectives of the research:

i) Visualise LV networks in a cost-effective manner (without extensive monitoring)

ii) Extract customers' energy patterns from the big data to granular levels (individual customer/day)

Due to the different characteristics of the load data from LV networks and individual customers, two different load profiling methods are developed respectively:

- A time-series load profiling is developed directly for LV networks. Firstly, instead of summing up the energy usage from different classes of customers, the method will for the first time directly cluster and classify LV substations. Secondly, in order to improve the visibility of LV networks with limited metering (cost), a set of *LV network templates* should be developed from limited but representative

metering samples, which can be spread to represent the load profiles of LV systems that are not monitored.

- A *spectral load profiling* method should be developed to process smart metering data based on their spectral features. As the individual load profiles are extremely volatile, uncertain and come with a massive amount of data, it is almost impossible to apply any techniques on time-series directly. It is promising to assess load profiling in the frequency domain, where the irregular load profiles can be characterised by the periodic spectral components, and big load data can be represented by a small number of spectral coefficients.

1.3 Research Challenges and Objectives

The state-of-art load profiling follows a two-stage clustering and classification process [19], which includes the following steps: i) cluster similar load shapes into groups. The TLP of each group will be determined by averaging load shape; ii) classifying an unknown customer into a proper customer group by recognising the customer's load profile pattern. However, it is difficult to directly apply the two-stage process to LV networks or individual customers due to a range of limitations:

1.3.1 Load Profiling for LV Networks

In the development of LV network templates, besides the practical constraints in limited monitoring devices, the main technical challenges are listed below, each with a potential objective *in italic*:

- **Lack of pre-knowledge and massive trial data delimit the performance of clustering analysis**

Unlike customers who can be usually pre-classified or at least macro-classified (e.g. residential, commercial or industry) before clustering [20, 21], LV substations cannot be pre-classified due to lack of available information, leading to difficulties in clustering and classification. Also, the big data from trial networks could challenge traditional clustering analysis. *One of the objectives would be to develop a method which can cluster data without any pre-knowledge or extra computational burden.*

- **Allocating an un-monitored substation into the right group (template) without any sample metering.**

In classification, customers are usually allocated to the most similar clusters by various pattern recognition techniques. It is difficult to implement on LV networks as no sampled data is available for un-monitored substations. Routinely the only available data is the fixed data, which are defined as network configuration and customer composition in this thesis. *The ideal solution would be able to classify LV substations solely based on fixed data.*

- **Reflecting both load magnitude and load shape of a LV substation**

In load profiling methods, load profiles are clustered based on their similarities on load shapes. Load data are normalised into per-unit load so that the developed clusters can reflect load shapes. However, the loading levels at LV substations, even of similar type, vary to a great extent [7]. It is inaccurate to represent their peaks by one average magnitude. *Thus it is critical to represent both load shapes and magnitudes for LV substations.*

1.3.2 Load Profiling for Individual Customers

In the investigation of load profiling for individual customers on individual days, the challenges brought up by smart metering data are:

- **Big Data on two dimensions**

Big data increases computational and storage burden. Two main drawbacks are identified in traditional feature selection techniques: i) they only reduce number of variables of each sample, but not the massive sample size on the other dimension; ii) Also they discard some sample points, which cannot be recovered, thus causing detailed information loss. *The objective is to develop a data compression method which can reduce data sizes on two dimensions while keeping detailed information intact.*

- **Volatility**

Most of the previous researches only concentrate on average load profiles because daily load profiles are extremely volatile. A sudden spike or a tiny time shift (communication delay) may lead to completely different clustering results. Different factors, such as magnitudes, overall trends and spikes will interfere with each other during the clustering. *A new clustering method is needed to cluster different factors without interferences.*

- **Uncertainty among individual days**

The same customer may have very different load profiles between days. Most researches [20, 22] used the averaged load profile over a time span (e.g. monthly) to prevent uncertainties in customer classification. However, averaged load profiles can be very different from individual ones, which form the non-convex sets. *The objective is to develop a load profiling method which can express the uncertainty between days in a probabilistic way.*

1.4 Research Contributions

Two novel load profiling methods are developed in this thesis: i) a novel three-stage load profiling method is proposed to visualise LV networks without extensive monitoring; ii) a spectral load profiling method which can visualise individual customers on individual days. The practical contributions of this work are as follows:

- Development of LV network templates, which significantly improves the visibility of LV networks without extensive monitoring cost. It reaches 87% accuracy in estimating the loading condition of un-monitored substation, which is validated by 6 Distribution Network Operators (DNOs) in the UK with over 200 substations in different areas.

- Development of spectral load profiling for smart metering analysis. It provides a promising solution to data compression, feature selection and customer classification. It can be efficiently used for DSR, load forecast, tariff design and settlement arrangements.

In the process, this study contributes innovative technique breakthroughs and fundamental scientific findings:

- Design a hybrid K-means and hierarchical clustering method for big data analysis on time-series. K-means clustering is highly efficient so as to release the computation burden from hierarchical clustering, which is in turn adopted to overcome the initialization issues of K-means.

- Classify un-monitored LV substations entirely based on routinely available fixed data. Multinomial Logistic Regression (MLR), a regression model aimed at predicting the outcome of a categorical dependent variable is adopted to classify highly mixed LV substations to distinguished clusters.

- Propose a Clusterwise Weighted Constrained Regression (CWCR) approach for LV substation peak demand estimation based on fixed substation information. A contribution factor is firstly brought up to address the contribution from a particular customer to the peak of different type of LV substations, which considerably enhances the estimation accuracy.

- Develop a two-dimension feature extraction method by spectral analysis. Big data are compressed on different dimensions: wavelets decomposition and reconstruction are used to reduce the number of coefficients (variables) describing each sample while a novel multi-resolution clustering (MRC) is proposed to work on the sample size reduction.

- Assess the performance of feature extraction of two decomposition techniques: DFT and DWT. Assessments are performed on load profiles from granularity to aggregation, showing DWT is more coherent with volatile and granular load profiles while DFT is more suitable to decompose smooth and aggregated ones.

- Develop a load profiling method for individual customers on individual days, MRC. It could effectively prevent inferences between different factors (e.g. magnitude, overall trend, spikes and etc.) by separating them to different resolution levels and clusters on each resolution independently. It also gives a probabilistic cluster membership instead of a deterministic one, addressing the uncertainty of cluster membership between days.

- More fundamentally, the work provides key contributions to big data analysis. As

shown in Figure 1-1, there is always a trade-off between big-data modelling errors and number of features and clusters. For feature extraction, excess features will add noises and redundancy into data while insufficient features will lose key information. For classification, a single cluster will mix up everything while too many clusters will lead to misclassification. Instead of treating them as two separate problems, this work aims to find the joint optimal number of features and clusters as the blue area shown in Figure 1-1.

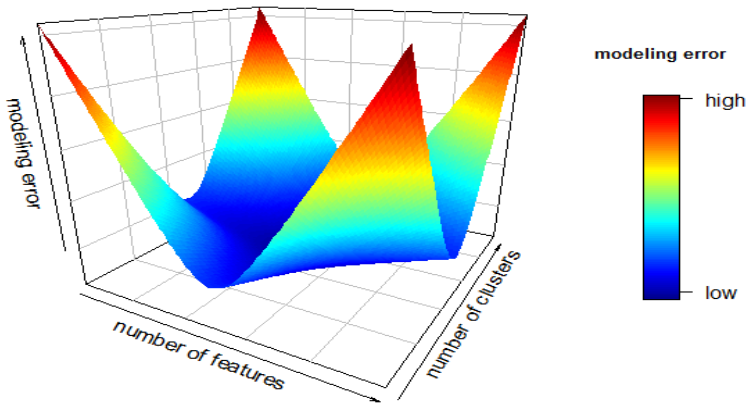

Figure 1-1 Sketch showing how anticipated modelling error varying with i) feature extraction; ii) classification

In summary, the research is conducted in two main streams. Firstly, in order to support LCN, a time-series load profiling method is proposed to visualise LV networks with limited cost and monitoring data. Secondly, for more efficient DSR, a spectral load profiling method is developed to provide accurate and granular load profiles at customer level. To overcome the significant limitations of existing techniques, this thesis will propose several novel techniques to develop new load profiling for LV networks and individual customer. Figure 1-2 presents a clear overview of the research.

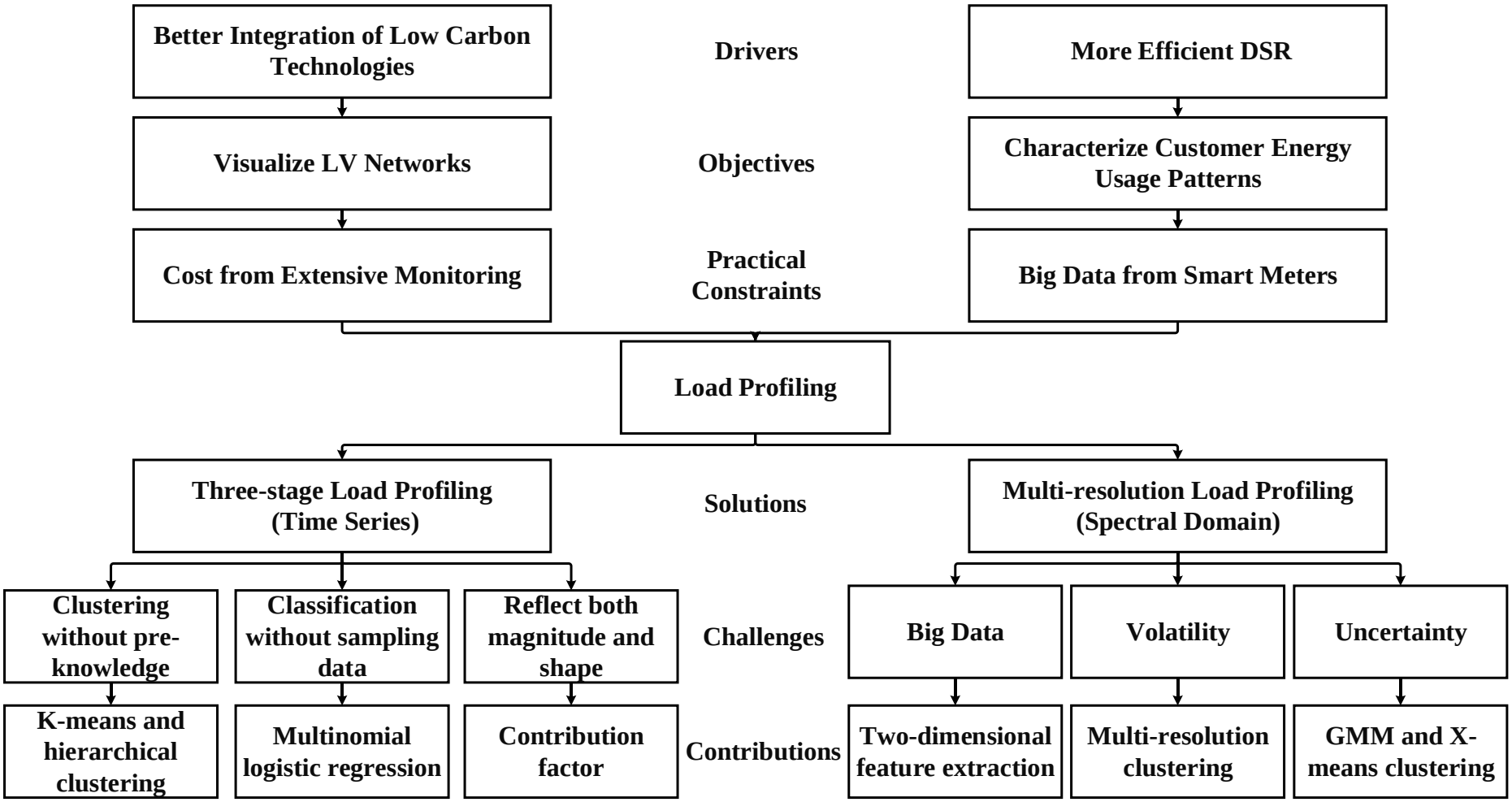

Figure 1-2 Overview of the research

1.5 Thesis Layout

The rest of the thesis is organised as follows:

Chapter two provides a comprehensive literature review of the load profiling methods in the UK and worldwide. The comparison will focus on two sections: engineering approach used by industry and reported approach in academia. For the development of new load profiling methods, techniques based on time-series analysis and spectral analysis are introduced and discussed with their own characteristics and applications.

Chapter three adopts the power synthesis approach to testify and analyse the representativeness of the existing load profiles in the new power environment. They are further tested with the recorded data taken from Dowlishford substation in Southwest England and several LV substations from South Wales. The corresponding errors are identified and the causes are analysed.

Chapter four proposes a novel three-stage network load profiling method. It uses real-time information monitored from selective representative areas to develop network templates. The three stages are: clustering, classification and scaling. This chapter will focus on the first two stages and chapter five will demonstrate the scaling. The method is demonstrated on a practical system in the UK under the umbrella of a smart grid trail project.

Chapter five follows previous chapter to develop the scaling stage. It proposes a novel contribution factor approach to predict diversified daily peak load of LV substations. The contribution factor is determined by a novel method - Clusterwise Weighted Constrained Regression (CWCR). It takes into account the contribution from different customer classes to substation peaks, respecting the natural difference in time and magnitude between LV substation peaks and the variance within the templates.

Chapter six assesses the performance of feature extraction of two decomposition techniques: discrete Fourier transforms (DFT) and discrete wavelet transforms

(DWT). The performance is evaluated by i) load characterisation: to decompose volatile load profiles consistently; ii) data compression: the trade-offs between the accuracy of the reconstructed profile and the degree of reduction in data sizes. Assessments are performed on load profiles from granularity to aggregation.

Chapter seven aims to classify customers from the extracted features from chapter six. It proposes a novel MRC method, which separates different load characteristics to different resolution levels and clusters on each resolution independently. The proposed method will be implemented on over 6369 smart metered customers from Ireland, and compared with traditional K-means clustering.

Chapter eight summarises the key findings from the research and the major contributions of the work.

Chapter nine provides some potential research topics in future work

Chapter 2

Review of Load Profiles

THIS chapter summarises a range of load profiles exercised by the UK and other countries. It also reviews different load profiling methods in the literature.

2.1 Introduction

For different applications, load profiles are developed by different process and techniques. Also, they could vary along with several factors in different countries. This chapter reviews three aspects of load profiles: the development of load profiles, the overall process and methodology applied, and the techniques deployed.

2.2 Development of Load Profiles

2.2.1 Definition of Load Profiles

In power system, the load varies with time and the supplier and network must respond to the customers' power demand. Therefore, time series load information is essential in different parts of power system activities including tariff design, load forecast, system planning and DSR [23-28].

Due to the diversity of customer type and extensiveness of distribution networks, it is impractical to collect load information of every customer continuously over time. The common approach is to assemble customers with similar load profiles into one group, and to study the typical load variations within each group at different time, day of week and seasons. For the LV networks, it is also a common practice to aggregate the customers load up to network level as an approximation.

Fundamentally, a load profile is a graph of the variation in the electrical load versus time. It visualises the load information and provides an aid to the tasks stated above. A load profile will vary according to customer type and other factors.

2.2.2 Factors of load profiles variation

- **Customer Type**

Load magnitudes and variation patterns of customers vary substantially. It is common to roughly divide customers into three groups: domestic, commercial and industry. They can be further classified into different sub-groups according to specific purposes. However, there are always variances within classes. It means even customers within the same group will still differ from each other, especially in terms

of peak and average consumption. It can be caused by internal factors including economic-social status and customer behaviours; as well as external factors such as weather [29].

- **Time and Climate**

Load varies all the time in a day naturally and it also changes through different days, seasons and years. Special days of year should also be considered separately. Load can be classified into four portions [30]: normal part, weather sensitive part, special event part and random part. Among them, weather including temperature, humidity, wind speed and sunset time, has the most significant influence on load variation due to the application of heating (cooling) and illumination.

- **Customer Composition and Network structures**

For LV networks, customer number and mix will strongly influence the load magnitude and shape. Although the load at LV substation is approximately the load aggregation of all customers served, different network structure could affect the accuracy by line and transformer losses. Network information including transformer type and rating, feeder number and length, load density and network structure are also able to indicate substation load to some extent.

2.2.3 Applications of Load Profiles

- **Settlement and Tariff Design**

Load profiles have been widely used for electricity settlement since the electricity market firstly opened up. In the absence of smart meters, customers are charged according to their estimated electricity bills, which are checked with the 6-month readings afterwards. However, as the market settlements are on a half-hourly basis, suppliers use load profiles to allocate the total consumptions for estimations and checking. Consequently, load profiles are used to guide the tariff designs. Early in 1990s, [31] raised a unit charging methodology in Finland based on load curve data. Several factors such as load factors, customer type and incomes are considered in tariff design with load profiles. [32] used load profiles to weight the contributions of

different customers to the system load and design tariffs according to the contributions. [33] raised two different approaches for price settlement based on load profiles: area model and category model. [25] uses load profiles to estimate the headroom of substations and margins left to DNOs for fixing dedicated tariffs to each customer class.

- **Load forecast and Network Planning**

Load profiles were used as an alternative tool for small area load forecasting due to the lack of metering data [34, 35]. They can be generally divided into two categories by data required and analysis methods. The first group of method is trending, which uses the historical load profiles to develop the trend of load growth for the future [36]. The other group of methods is multivariate methods involving work with customer type, demographic, economic and network data. Techniques such as time series, regression and [37, 38] neural networks are used to formulate the relationship between load profiles and various factors [39, 40]. Load profiles are also widely used to estimate the loading level for network planning and assets capacity design.

- **Other Applications**

[23] uses load profiles in DSR in order to let customers participate directly in the electricity market. Load profiles notify customers of the necessary demand scheduling and assist them in the way of demand response to demand curtailment period and prices. Load profiles are also used in many other fields, like network state estimation [41] and distribution transformer loss-of life evaluation [42].

2.2.4 Load Profiles across the World

- **The UK**

In the UK, the Electricity Association has studied loads in England and set about a program [43] of analyses in order to define the number and type of profiles to be used in settlement. Customers with power demand above 100 kW are equipped with half-hourly meters. The rest of the customers are roughly divided into 8 classes and eight generic profile classes were developed which can represent large populations of

similar customers. The eight profile classes and their characteristics are given in Table 2-1 [44].

Table 2-1 Load Classes and Their Descriptions

Profile No.	Description
Class 1	Domestic Unrestricted Customers
Class 2	Domestic Economy 7 Customers
Class 3	Non-Domestic Unrestricted Customers
Class 4	Non-Domestic Economy 7 Customers
Class 5	Non-Domestic Maximum Demand (MD) Customers with a Peak Load Factor (LF) of less than 20%
Class 6	Non-Domestic Maximum Demand Customers with a Peak, Load Factor between 20% and 30%
Class 7	Non-Domestic Maximum Demand Customers with a Peak, Load Factor between 30% and 40%
Class 8	Non-Domestic Maximum Demand Customers with a Peak, Load Factor over 40%

With the deviation of power demand in different days (weekday and weekend) and different seasons (spring, summer, high summer, autumn and winter), profiles within each class are sub-classified. Taking Winter Wednesday as an example, eight-class profiles for this season and day type are depicted in Figure 2-1 to 2-3.

Class 1 and Class 2 describe domestic load profiles installed with normal and economy 7 types of electricity meters. In contrast, Class 3 and Class 4 represent non-domestic load profiles with the same type of meters. Classes 5, 6, 7, and 8 are used to divide different non-domestic customers by maximum demand and peak load factor. Additionally, Class 00 represents large consumers installed with half-hourly metering. Large consumers indicate industrial users who are directly connected to 11kV.

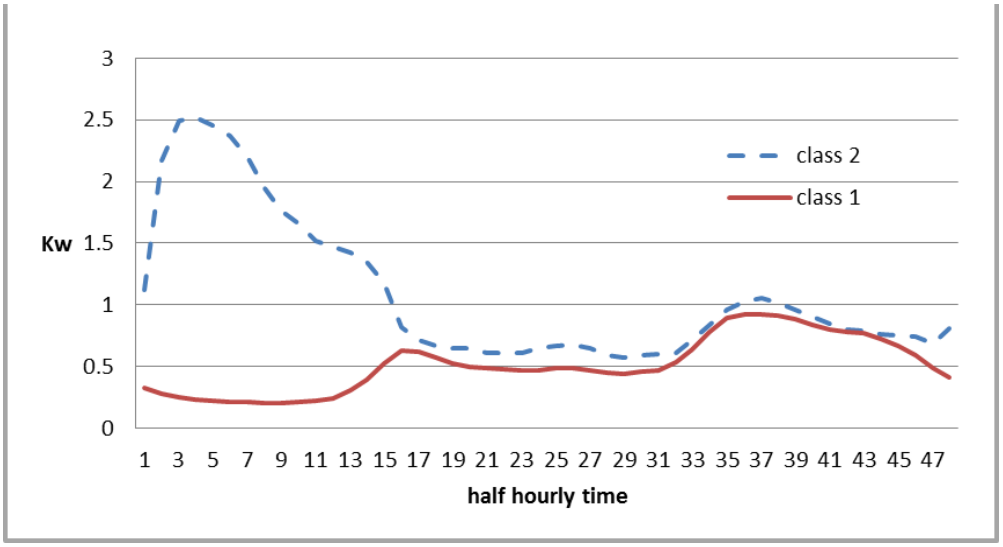

Figure 2-1 Domestic profile classes for winter Wednesday

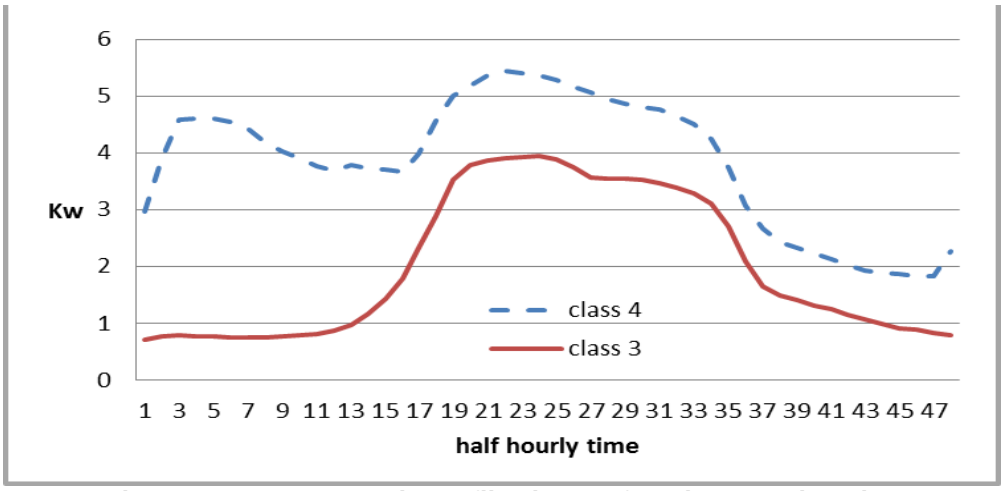

Figure 2-2 Non-Domestic profile classes for winter Wednesday

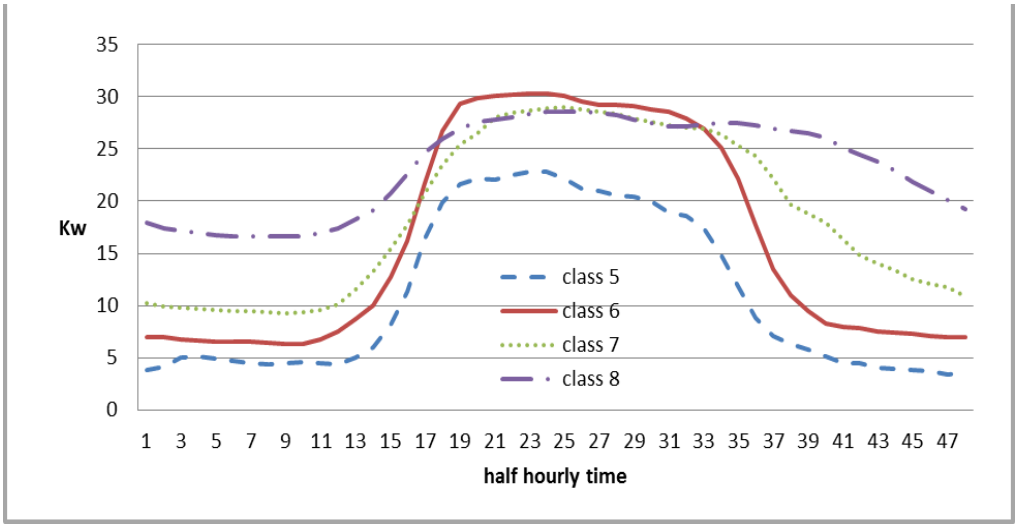

Figure 2-3 Non-Domestic Maximum Demand profile classes for winter Wednesday

- **Norway**

In 1994, Norwegian Water Resources and Energy Administration developed the standard load profiles for unmetered customers in order to encourage them participate in the power market [45]. The load profiles were generally developed under the area model, which treats all the unmetered customers in one area (served by same substation) as a whole. Total load profiles of these customers are derived by deducting metered large customers and losses from the total substation load.

- **Finland**

Finland has a long history of load research and has done vast studies in different aspects. In Finland, the category model is used to develop load profiles. Only customers with 3*63A main fuses or above are metered hourly, the rest of customers are divided into four classes: 1. Households without electric heating. 2. Households with electric heating. 3. Other customers above 3*35A fuses. 4. Other customers below 3*35A fuses. The classification is made beforehand based on fixed information.

2.3 Load Profiling Methods

2.3.1 Customer Classification and Load Shape Clustering

The primary goal of load profiling is to arrange customers into groups according to their shared characteristics on load profiles. There are two common stages in this process: clustering and classification. The difference, in general, is that classification tries to allocate a new object into a set of pre-defined classes while clustering tries to group objects without pre-labels and to discover the relationship between objects. In the context of machine learning, classification is supervised learning and clustering is unsupervised learning.

- **One-stage Pre-classification**

This group of methods usually pre-classify customers based on their fixed data. The concept of fixed data is compared with that of metering data. Fixed data are those

barely vary with time, such as social-eco information of customers and configuration of networks. The feature of these methods is that they all start with developing a classification rule based on fixed data [46-49]. All customers are assigned into predefined classes according to fixed information like tariff types, certain appliances (e.g. electrical heating) or the nature of business. Samples will be usually selected from each of the classes and monitored through periods. The TLPs will be derived as the average within each group and within each pre-defined time period (e.g. seasonally). For new customers, as long as the fixed information is available, they can be easily classified according to the pre-defined classification rule. Figure 2-4 clearly demonstrates the process of this type of method.

The advantage of these methods is that they provide clear and accurate classification rule, which can be easily applied to new customers. However, the disadvantages also exist: i) customers within same classes may have very different load profiles. Customers having similar fixed data do not always share similar load profiles; ii) in turn, customers in different classes may have similar load profiles.

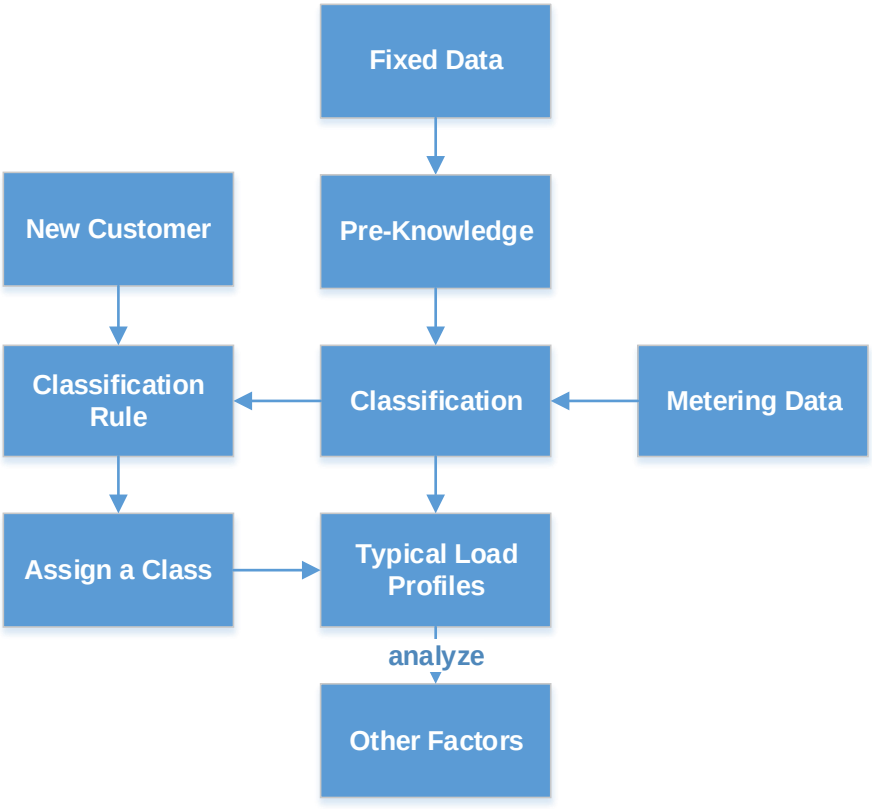

Figure 2-4 Flow charts of classification methods by fixed data

- **Two-stage clustering and classification**

As shown in Figure 2-5, the second groups of methods, on the other hand, does not predefine substation classification, but rely on un-supervised clustering analysis to find the underlying similarity between load profile shapes [50-53]. Various clustering methods have been used, including self-organised maps, K-means clustering, fuzzy C-means, etc. to find substations with similar load shapes and group them into the same clusters [21, 41, 54]. The second stage is to classify new customers into proper group. A piece of sampled data would be usually required from the new customer. By pattern recognition, the sample will match with the group which has the most similar TLP.

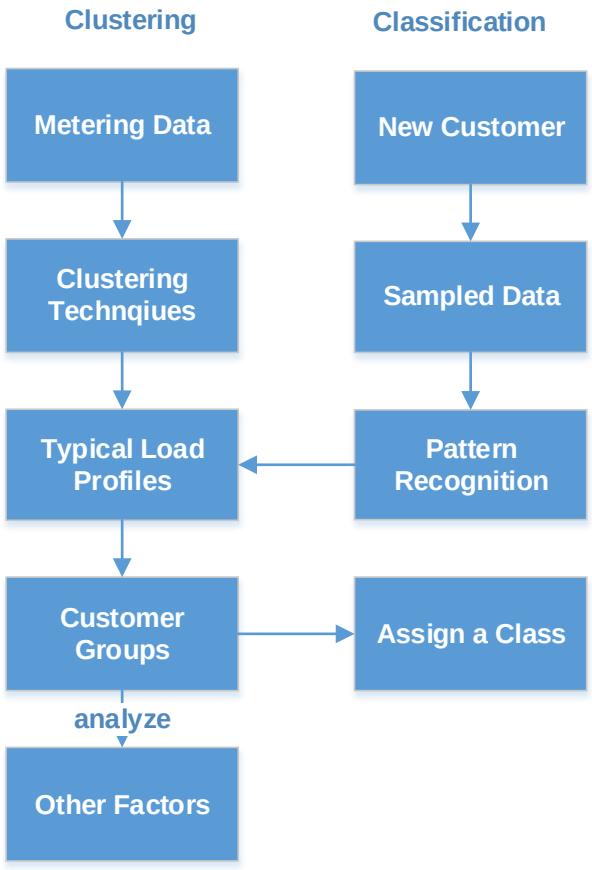

Figure 2-5 Flow charts of classification methods by fixed data

The obvious advantage of these methods is that they ensure the similarity within each group. But the disadvantages are: i) sampled data are sometimes unavailable, especially for LV networks; ii) customers (substations) with similar load profiles may have very different fixed information, which means that it can be difficult to define each group or to investigate the customer behaviour behind it.

2.3.2 Load Magnitude Scaling

The units of the data used as input for the clustering to a large extent determine the nature of the clusters. As the magnitudes of different customers (substations) vary substantially, the direct use of the measurements in kW will produce clusters that only reflect the magnitude of loads but not their shapes. Therefore, normalization of the data is usually conducted to convert data into per unit load. The clusters developed can reflect the load patterns/variations within a day; however, the magnitude information would be missing. A scaling is needed to reflect the magnitude of the normalised TLPs. For mass customers, previous studies usually take the average magnitude of the group. For LV networks, two main methods are reported to estimate the loading level of a LV substation.

- **Engineering P-Q method**

In the UK, the industry traditionally uses annual energy consumption to estimate peak load for LV network design [15, 55]. The basic idea is by using statistical methods to convert the annual consumption of LV substations to winter mean peak power demand. The targeted peak demand is calculated by adding a certain level of standard deviation onto the mean peak demand. This method is based on the following three assumptions:

i) The mean demand of a LV substation is proportional to the annual consumption

$$\overline{L} = N \cdot C \cdot \Psi \tag{2-1}$$

where, $\overline{L}$ is the mean demand of a class of similar consumers, N is the number of customers in the class, C is the mean annual consumption of each customer in the class, and Ψ is the factor converting consumption (kWh) to mean demand (kW), known as Demand Estimation Coefficient (DEC) and varying for different classes of customers.

ii) Because the power demand follows normal distribution so that the peak demand can be expressed as the mean demand added by certain standard deviation

$$L_p = \bar{L} + \beta \cdot \sigma = N \cdot C \cdot \Psi + \beta \cdot \sigma \tag{2-2}$$

where, L_p is the peak load of a class of similar consumers, σ is the standard deviation of $\bar{L}$,and β decides the number of standard deviation added to the mean load [14].

iii) It assumes the peak load occurs in the coldest winter night of a year.

- **Conversion and Diversity Factor (C-D) Method**

Another method adopts kWh-to-peak-kW Conversion factors (C factor) and Diversity factors (D factor) to estimate LV substation peak load based on customer billing cycle kWh consumption. It is also known as C-D method [56].

Considering the daily load profile as a discrete-time series over an interval [0, T] divided into *n* subintervals, the load profile of the m^{th} LV substation is defined as: $L_m = [L_m(t_1), L_m(t_2), \cdots, L_m(t_n)]$. The aggregated daily load of class *i* customers is $S_i = [S_i(t_1), S_i(t_2), \cdots, S_i(t_n)]$. The daily peak of LV substations, $\max(L_m) = L_m(t_q)$, can be derived by (2-3), where the C and D factors are defined in (2-4) and (2-5).

$$L_m(t_q) = \sum_{i=1}^{I} E_i \times \frac{C_i}{D_{in}} \tag{2-3}$$

$$D_{in} = \frac{\sum_{i=1}^{n} P_{ij}}{S_i(t_p)}, \quad \text{where } S_i(t_p) = \max(S_i) \tag{2-4}$$

$$C_i = \frac{\sum_{j=1}^{n} P_{ij}}{E_i} \tag{2-5}$$

where, $L_m(t_q)$ is the estimated peak load of a substation with the total *I* classes of customers; $S_i(t_p)$ is the estimated aggregated peak load of class *i* customers; E_i is the total energy consumption of class *i* customers served by the substation; D_{in} is the diversity factor within all *n* customers in class *i*; P_{ij} is the peak load of an individual customer *j* in class *i*; C_i is the conversion factor for customer class *i*.

2.4 Chapter Summary

The chapter gives an overview of the load profiles in terms of definition, characteristics and applications. It introduces the development of load profiles in the UK and other countries briefly. The existing load profiling methods are reviewed and summarised into two categories: i) one-stage classification, and ii) two stage clustering and classification. For the magnitude of load, existing P-Q and C-D methods are presented.

The main limitations of existing load profiling methods are: i) absence of a direct way to visualise the LV networks; ii) lack of capability to reflect both load magnitude and shape; iii) the inaccuracy at granular level for individual customers.

Chapter 3

Evaluation of Engineering Load Profiles in the UK

THIS chapter conducts assessments of the load profiles currently used by the UK power industry. The tests are implemented on different voltage levels. The results show significant errors especially at more granular levels.

3.1 Introduction

As introduced in Chapter 2, 8 TLPs currently used by the UK electricity industry were developed in 1990s [12]. Considering the fast development in the power industry and LCTs during the last 20 years, the customers' energy usage habits have probably changed to some extent where the current load profiles can no longer match them. Therefore, a test of the accuracy and representativeness of the current load profiles is necessary. For this thesis, two aspects of load profile accuracy are examined:

i) representativeness of individual customers' load profiles;

ii) ability to reflect the aggregated energy for distribution networks

This chapter will mainly focus on test ii) as it involves an indirect bottom-up approach to estimate substations' load profile from the customers'. A power synthesis method is demonstrated in this chapter to examine the repetitiveness and relevance of the current eight load profiles. The work is conducted on a typical British High Voltage (HV Distribution -11kV and 6.6kV in the UK) power substation named Dowlishford taken from the Western Power Distribution (WPD) network in Southwest England and several LV substations in South Wales. For individual customers, the TLPs are also compared with the smart metering data from the Irish smart meter trial project [57]. There are 6369 customers with half-hourly demand recorded over one and a half years (2009-2011). The detailed assessment of individual customers will be demonstrated in Chapter 6.

3.2 Power synthesis method for Load Profile Test

Due to the limited monitoring devices on distribution networks, especially the LV networks, DNOs usually estimate their load profiles by indirect power synthesis, which aggregates the load profiles of all connected customers. This chapter firstly aims to assess the accuracy of this method by comparing their aggregation with metered network load profiles. The main idea is to evaluate whether the synthesised TLPs from customers can reflect the real loading condition of distribution networks. It includes the following steps: i) determine the customer size of the area served by the substation being tested; ii) classify the customers into one of the eight customer

profile classes; iii) aggregate the customer energy usage patterns within each class to obtain the class profiles; iv) aggregate all eight classes profiles to obtain the estimated substation network profile; and v) compare the estimated profile with the metered network profile from the substation to verify the accuracy and applicability of the currently used eight load profiles. The detailed implementation steps of the proposed approach are depicted in Figure 3-1.

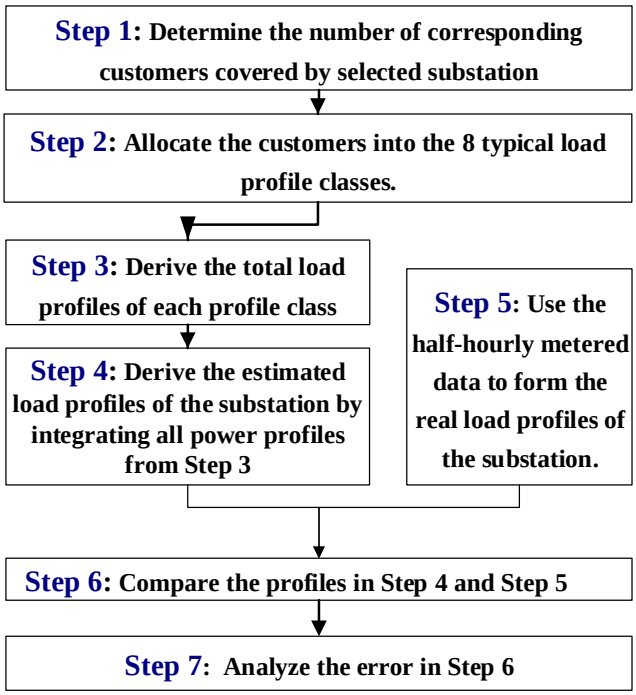

Figure 3-1 Overall process of power synthesis method

3.3 Customer Classification

3.3.1 Customer size

The work is firstly demonstrated on a typical British distribution power substation named Dowlishford The substation mainly serves the areas of Ilminster, Ilton, Shepton Beauchamp, Neroche, and Barrington. The number of households in year 2001 of Ilminster and Neroche can be found in Neighbourhood Statistics [58]; however, the household number in Ilton, Shepton Beauchamp and Barrington is not available .

For the number of households in Ilminster and Neroche : The number in 2011 is not promptly available, but could be obtained by assuming that the population growth

obeys exponential pattern, given as

$$N(t_K) = N(t_0) \times (1 + R)^K \tag{3-1}$$

Where, $N(t_k)$ is the number of households in the year t_k, $N(t_0)$ is the number of households in an initial year t_0, R represents compound annual growth rate of population, and K stands for the years between t_k and t_0.

For the household number in Ilton, Shepton Beauchamp and Barrington: The numbers of households in these areas were estimated based on the number of population [59] by (3-2). The numbers of households are also listed in Table 3-1.

$$n_h = \frac{N_h}{N_p} \times n_p \tag{3-2}$$

where n_h is the number of households in a certain area, N_h is the number of households in Southwest England, N_p is the population in Southwest England and n_p is the population in a certain area.

Table 3-1 Number of households served by dowlishford substation in 2011

Town/village	Number of households	
	2001	2011
Ilminster	2209	2488
Ilton	***	450
Barrington	***	226
Shepton beauchamp	***	350
Neroche	1002	1128
Total	***	4642

The average annual growth rate of population in the UK is approximately chosen as 1.2% per year [60], and thus the number of households in each area 2011 can be determined, listed in Table 3-1. Once the household numbers in each area served by

the substation are obtained, every household needs to be clustered into one of the eight existing profile classes.

3.3.2 Classify Customers into Eight Classes

The numbers of domestic and non-domestic consumers in Southwest England are found from regional and national authority electricity consumption statistics and the ratio between domestic consumers and non-domestic consumers is around 10:1. Table 3-1 indicates that there are 4642 domestic customers served by Dowlishford substation. It is calculated that the number of non-domestic consumers at this area is approximately 464 [61]. The number of non-domestic consumers is determined by

$$n_{non} = \frac{N_{non}}{N_d} \times n_d \tag{3-3}$$

Where, n_{non} is the number of non-domestic consumers in the study area; n_d is the number of domestic consumers in the study area; *Nnon* is the number of non-domestic consumers in Southwest England; *Nd* is number of domestic consumers in Southwest England. As stated before, domestic customer classes include class 1 and class 2. Further specified number of customers in each domestic profile class is determined by (3-4)

$$n_{dj} = \sum_{i=1}^{k} \frac{C_{di}}{T_{di}} \times \frac{C_{dj}}{T_{dj}} \times n_d \tag{3-4}$$

Where, n_{dj} is the number of customers per profile class *j*; C_{di} is the total annual power consumption of class *i* customers in Southwest England; C_{dj} is the total annual power consumption of class *j* customers in Southwest England; T_{dj} is the typical annual power consumption of a single class *j* customer; T_{di} is the typical annual power consumption of a single class *j* customer; n_d is the number of customers per profile class *j*; *k* is the total number of domestic profile classes.

The numbers of non-domestic customers per rest profile classes are determined by the same method. The annual consumption for each customer group is obtained from the Common Distribution Charging Methodology for the Southwest area. The relative

ratio between the customer classes and retained the same ratio for this area is subsequently calculated [62]. The numbers of customers (households) within each of the eight typical customer classes are presented in Table 3-2.

Table 3-2 Households numbers of eight typical load profiles

Profile class	Number of MPANs
class 1	3,212
class 2	1,430
Total (Domestic)	4,642
class 3	330
class 4	88
class 5	22
class 6	12
class 7	8
class 8	4
Total(Non-domestic)	464

3.4 Estimation of Substation Profiles

The basic idea used in this test is that the load profile at a distribution substation is equal to the aggregated profiles of all customers it supports, assuming that the losses on transformers and distribution lines can be neglected. An example describing the estimated substation profile for winter Wednesday is shown in Figure 3-2.

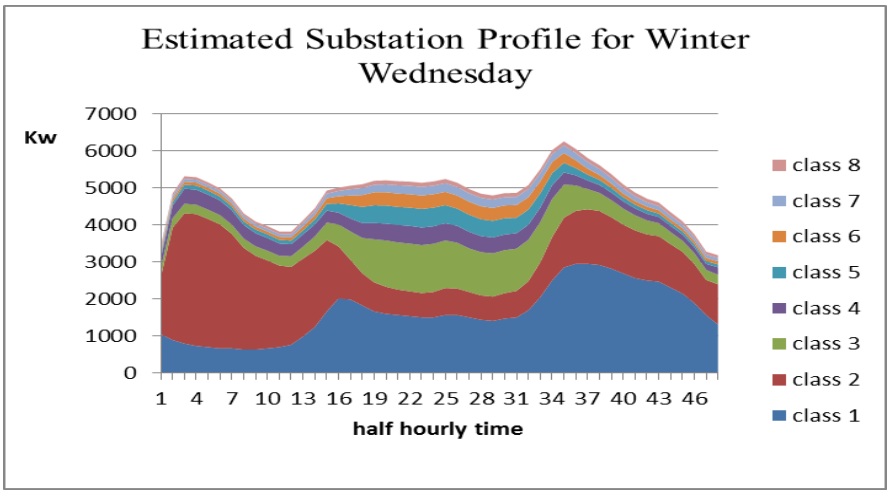

Figure 3-2 Estimated substation profile for winter Wednesday

As can be seen from Figure 3-2, class 1 customers, who are unrestricted domestic customers, contribute most to the total power consumption from morning to midnight; on the other hand, class 2 customers, who are domestic economy 7 customers, dominate the power usage after midnight. This is explained by the fact that economy 7 customers, who have 2 price rates, use some electrical appliances after midnight to take advantage of the lower rate.

However, the non-domestic customers have relatively lower power consumption through a day. The reason might be the area served by this substation contains more domestic customers than non-domestic customers compared with metropolitan areas.

3.5 Derivation of Metered Substation Profiles and Comparisons

3.5.1 Substation power profiles

To test the accuracy of estimation load profiles above, real power consumption patterns at Dowlishford substation are used as reference to be compared with. Real-time voltage and current data have been recorded every half hour through a whole year. Figure 3-3 shows the configuration of Dowlishford substation.

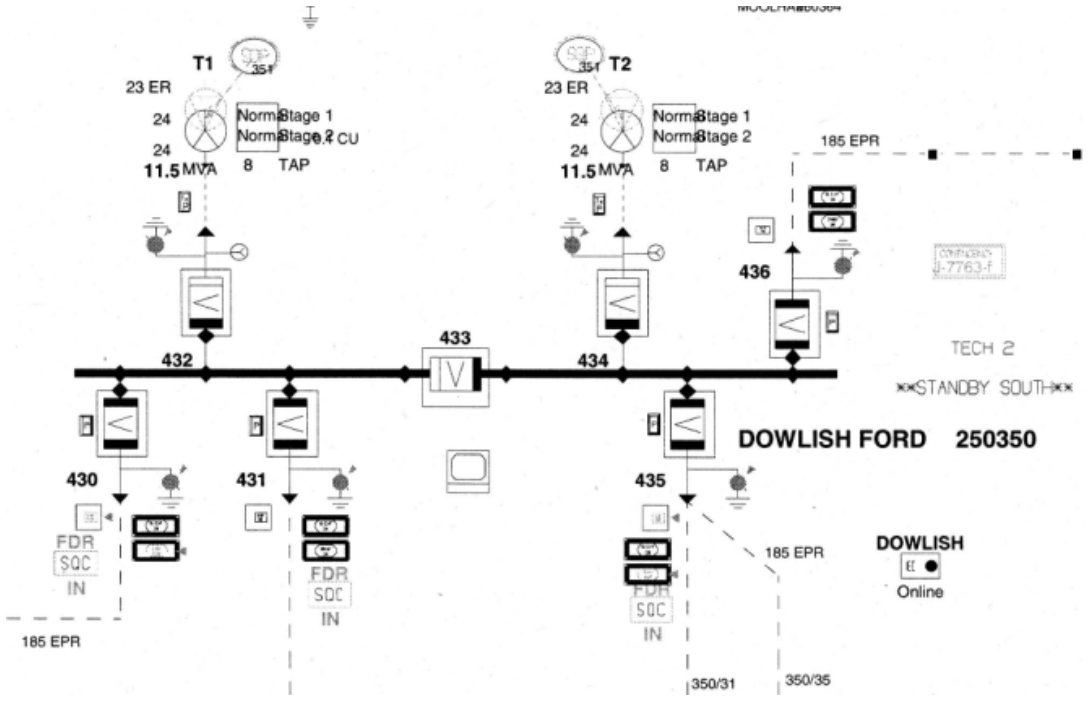

Figure 3-3 System map of Dowlishford substation

As can be seen from Figure 3-3, there are 2 main transformers T1 and T2 at Dowlishford substation, which are protected by two circuit breakers (CB), CB432 and CB434. Voltages and currents on both CBs are metered and recorded every half hour. The data of 2010 is available from WPD. The recorded voltages and currents are utilised to calculated the power flow on the two CBs with (3-5)

$$P = \sqrt{3} \times U \times I \tag{3-5}$$

Where, P is the real power, U is the line voltage, and I is the line current.

3.5.2 Comparisons

The comparison between the estimated load profiles derived in Section V and the metered load profiles from Dowlishford substation calculated above is conducted here. Taking two extreme cases, the analysis only focuses on the comparison between the estimated and metered profiles of two typical Wednesdays in winter and summer scenarios, given in Figure 3-4 and 3-5.

As seen, the shapes of the estimated profiles are very similar to those of metered data in both figures. This indicates that the current customer classes are generally reasonable and the eight benchmark profiles are basically able to reflect the energy usage patterns and habits of different classes of customers.

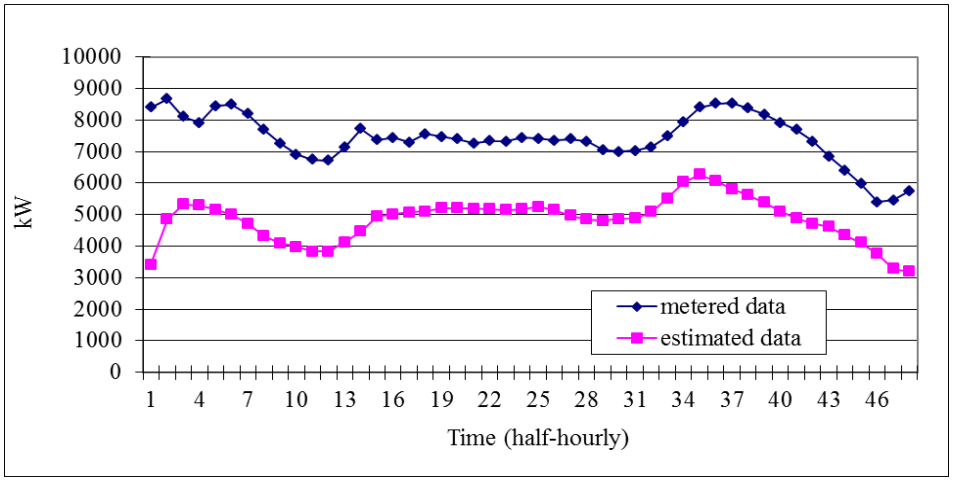

Figure 3-4 Comparison between metered and estimated data for winter Wednesdays

However, the magnitudes of the estimated data are smaller than the metered data in both seasons and it is noticeable that the differences between the estimated data and

the metered data are consistent through a day. To analyse this regular difference, it is calculated that the real metered electricity consumptions are overall 1.5 times higher than the estimated. The reason is probably that with the development of economy and the trend of electrification, the consumption of electricity has been increasing for decades [63]. Due to the increased use of electrical appliances and consumer electronics in the home, the domestic electricity consumption increased by 59% from 1970-2009 [64]. This increasing rate is very close to the 50% difference in this study. Therefore the differences between the estimated data and the metered data probably come from annual electricity consumption increase.

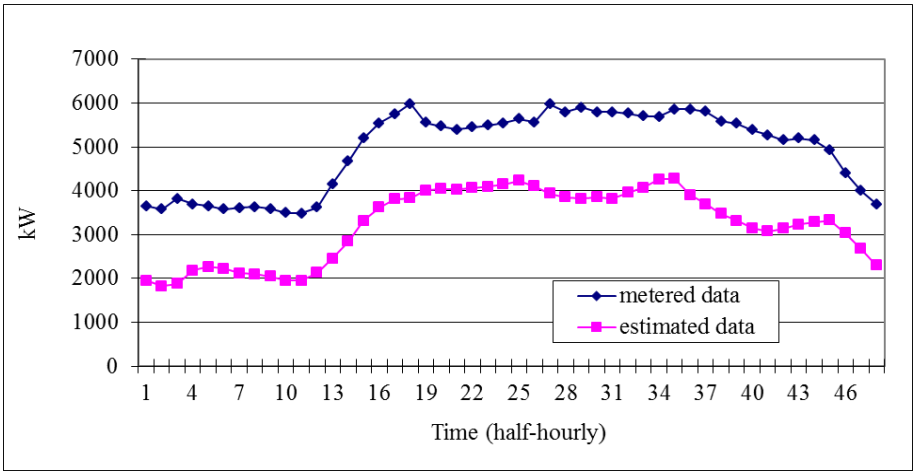

Figure 3-5 Comparison between metered and estimated data for summer Wednesdays

Additionally, Figure 3-4 shows that the metered electricity consumption during winter midnight is substantially higher than estimated and the patterns differ to some extent. The metered consumption at winter midnight is approximately 2 times higher than estimated, which is higher than the overall 1.5 times. This might be caused by several reasons. Firstly, the wide use of electricity heating boosts the domestic power consumption in winter nights. Secondly, the changes of night rate of economy 7 customers in recent years may stimulate the increase of class 2 (economy 7) customers and their electricity consumption during night.

3.5.3 Other Potential Causes

The patterns of estimated profiles conform well to those metered, but the magnitudes differ greatly throughout the studied year and the differences are quite similar. Although part B has summarised the main reasons, some other potential causes may

also slightly contribute to this difference:

i) Class 00 customers: it was assumed that the profile class 00 customers contribute little to the substation profile. However, due to the lack of information on class 00 customers, it is possible that they may influence the substation profile to some extent. As mentioned in VI B, the differences between the estimated data and the metered data are about 50%. On the other hand, the difference can be treated as a constant value and it is measured around 2000kW. It is possibly that class 00 customers have an impact on the substation profile because the typical profile of class 00 customers is flat and constant.

ii) Ratio between profile classes: based on the fact of high consumption during winter night and that class 2 and class 4 customers, who are economy 7 customers, contribute the largest portion of the electricity consumption during night, it is suspected that the number of profile class 2 and 4 customers in the town of Ilminster might be higher than the average in southwest. Thus the estimated numbers of class 2 and class 4 customers might be smaller than actual.

iii) Losses on line: from the substation to the customers, electricity is transmitted via distribution lines and stepped down to standard voltage level trough service transformers. During this process, some of the electricity is lost on lines and transformers. Therefore the power metered at substation is actually the load plus the losses.

3.5.4 Test on Low Voltage Level and Individual Customers

As the test goes to less aggregated level, the results are quite different from higher voltage level because LV substations and individual customers have more volatile and irregular load profiles.

Figure 3-6 shows the comparison between proxy profiles based on the typical eight load profiles and two LV substations in South Wales on 11th July. In the figure, two substations are represented by their substation number 531832 (black) and 521842 (red). These two substations share very similar customer number, customer composition and network structure. Using the same method, a proxy profile based on

TLPs can be derived (green) to estimate the load profiles of the two substations. It can be seen from the three load profiles, although two substations real load profiles are very similar, our estimated profile is actually different from them in terms of shape and magnitude.

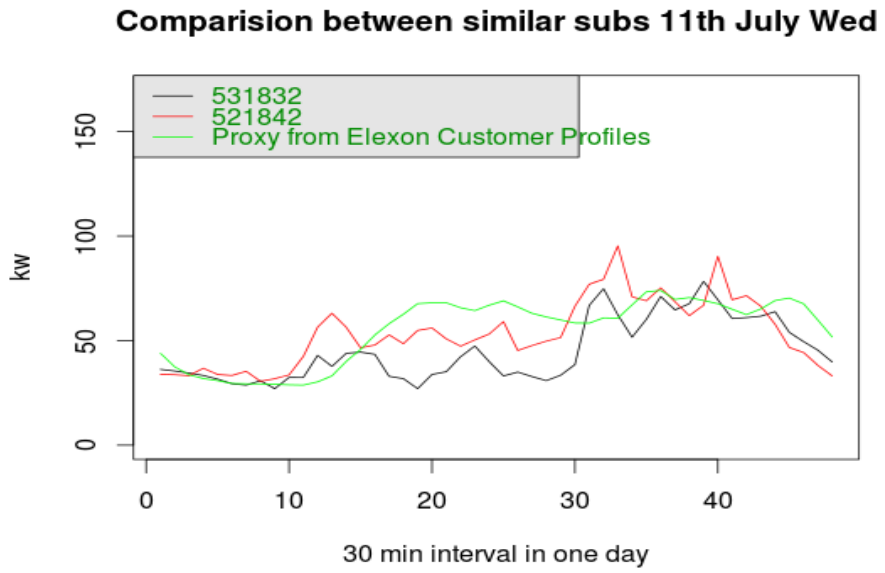

Figure 3-6 Comparison between similar LV substations on 11th July

Compared with the results in Dowlishford substation, which is at 11kV, the results at LV substations are much less representative. The reason is mainly that the eight load profiles developed by Elexon are used to represent average customer load profile, which is naturally more accurate when using under a large number of populations. However, LV substations are less aggregated, which means there are less customers connected under a LV substation than a HV/MV substation. The total over 30 million domestic households in the UK are now represented by only 2 types of load profiles (class 1 and 2). It leads to a huge variance within each class as expected. As the traditional TLPs are initially designed to indicate high-level aggregation, they are unable to pass any load information of individual customers. the Figure 3-7 demonstrates the comparison between the TLP (class 1) and real load profiles from smart meters. In the figure, the red line is the class 1 TLP in summer weekday scenario while the six grey lines are six random domestic customers on random summer weekdays. For any individual household, the load profiles are extremely volatile and irregular. Clearly, the traditional TLP cannot neither reflect the overall magnitude nor capture the spikes.

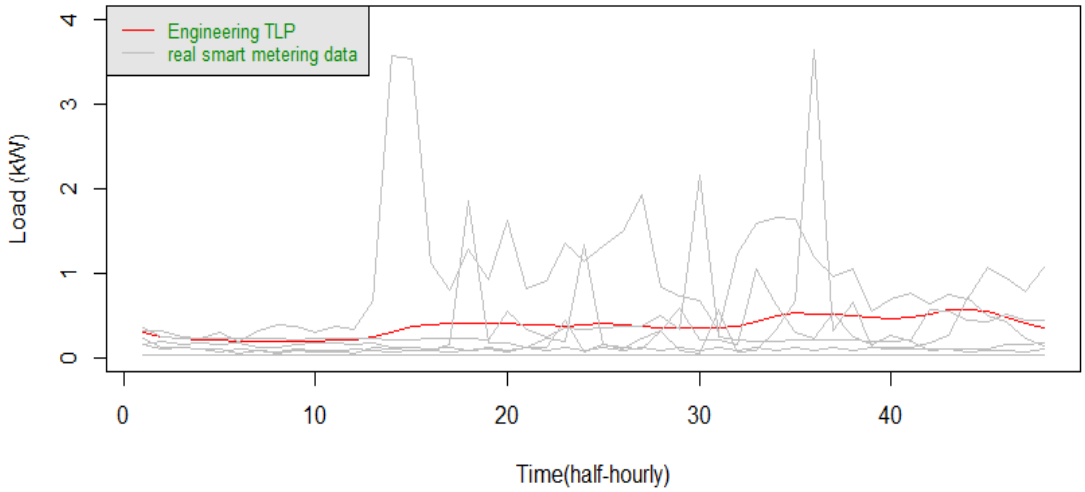

Figure 3-7 Comparison between similar LV substations on 11th July

3.6 Chapter Summary

It is found that the estimated profiles which are produced by integrating the existing eight profiles conform well in shape to the real load profile metered at HV level but differ in magnitude. For LV networks, it is found that the traditional load profiles are mostly unable to mimic the load profiles, which are less aggregated. Further, they cannot reflect any of the more volatile and irregular load profiles of individual customers.

Based on these factors, it can be concluded that

i) Traditional indirect bottom-up approach cannot accurately express the load profiles for LV networks;

ii) the current eight classes of customers load profiles can no longer reflect the power consumption pattern of today's customers

It is therefore critical now to develop new load profiling methods to accurately describe the loading condition of LV networks and the energy usage of individual customers.

Chapter 4

Time-series Load Profiling for LV Networks: Clustering and Classification

THIS chapter proposes a novel three-stage load profiling method for LV networks. It develops a set of LV network templates, which can directly visualise the LV networks without extensive monitoring.

4.1 Introduction

Chapter 3 has tested the traditional indirect load profiling method for LV networks, which aggregates the TLPs of all connected customers. However, the representativeness was compromised by i) the inaccuracy of customer load profiles; ii) limitations on information and iii) the neglect of network characteristics. In order to visualise the LV networks, the most direct way would be wide installation of monitors, which can provide real-time loading information. However, such approach can be prohibitively expensive.

This chapter proposes a novel three-stage network load profiling method. It is the first attempt at load profiling LV substations directly. The basic idea is to use small but selective monitoring samples to develop a set of LV network templates, which can be used as benchmarks to represent the remaining unmonitored ones. The templates can significantly improve LV network visibility without extensive monitoring and integrate LCTs in a cost-effective manner. The three stages are: clustering, classification and scaling. In the clustering stage, hierarchical clustering and K-means are used to cluster substations into groups based on the variations over time of the monitored load profiles. The classification tool designed with MLR maps an unmonitored LV substation into the most probable templates by using routinely available fixed data. Finally, Clusterwise Weighted Constrained Regression is employed to estimate peak load for individual LV substations and the developed templates.

The three-stage profiling is demonstrated on a practical system in the UK under the umbrella of a smart grid trial project. 10 LV templates are developed by using the metered data from 800 monitored LV substations. A series of industrial validation has indicated that the three-stage process can achieve superior accuracy than traditional methods. This chapter will firstly introduce clustering and classification. The scaling (peak estimation) process will be introduced in Chapter 5.

4.2 Problem and Proposed Solution Statement

A major change in distribution networks in the coming decades is the substantial increase in LCTs, driven by the governmental ambition in reducing greenhouse gas

emissions and improving supply efficiency [1, 65]. These technologies, such as EVs, HPs, photovoltaic (PV), and other smart appliances [66, 67], will be largely connected to LV distribution networks. They thus will require DNOs to understand the capabilities of the existing LV networks so as to assess whether they can be accommodated.

Unfortunately, DNOs currently have very limited visibility and knowledge of LV networks in terms of real-time network utilisation. They largely rely on fixed network information, such as customer number, type and electricity use behaviour, to estimate annual or daily peak demand of a LV substation by aggregating typical customer load profiles served by the substation [5, 15]. Although peak load estimation based on fixed information is very economical, it is highly inaccurate [2] particularly if DNOs are interested in daily loading conditions. The reason is that the load profiles at LV substations are volatile, irregular and noisy compared to those at HV substation level.

One approach for visualising LV networks is to install monitoring devices at every single substation, but it is prohibitively expensive. In the UK, there are more than 900,000 LV substations and it would cost over £2 billion to install metering and data acquisition equipment, and do daily data management. An economical alternative is network load profiling, which identifies TLPs from a limited samples/areas to represent the load profiles of LV systems that are not monitored.

Load profiling has been widely used at household level for customer classifying and profiling. Most of the early work in [32, 47, 48, 68] pre-defines customers' classification according to pre-knowledge, such as customer type and characteristics, and derives a TLP for each class based on sample metered data. The limitations of these methods are misclassification, potentially leading to dissimilarity of load profiles within a pre-defined customer class. In order to overcome the limitations, more recent studies in [20, 69, 70] tend to [21, 41, 71, 72] follow a two-stage clustering and classification process as proposed in [19]. It includes the following steps: i) clustering similar load shapes (normalised load profiles in [0, 1] range) into groups. The TLP of each group is determined by averaging load shape and weighted average magnitude within the group; ii) classifying an unknown customer into a proper customer group by recognising the customer's load profile pattern.

At LV substation level, there is not yet an established technique for load profiling, particularly by using limited metered information. The direct application of the two-stage clustering and classification process for LV networks has a number of limitations: i) unlike customers who can be usually pre-classified or at least macro-classified (e.g. residential, commercial or industry) before clustering, LV substations cannot be pre-classified due to lack of available information, leading to difficulties in clustering and classification; ii) in classification, customers are usually allocated to the clusters with the most similar patterns. It is difficult to implement on LV networks as no sampled data is available for un-monitored substations. Routinely available data is only fixed data, such as network configuration and customer composition; iii) in load profiling methods, magnitude information has rarely been considered and they simply take weighted average magnitude within a customer group to estimate the peak. The loading levels at LV substations, even of similar type, vary to a great extent [7], inaccurate to represent their peaks by one average magnitude.

This study proposes a three-stage load profiling method for visualising LV substations by using limited but representative metered real-time load data at LV substations. The work consists of the following three steps:

- Clustering: to group substations according to similarities in load shapes, where normalised TLPs are developed for each substation cluster, defined as normalised templates.

- Classification: to characterise the relationship between templates and fixed data of substations so as to map unmonitored substations to an appropriate template.

- Scaling: to estimate the daily peak load for individual substations in order to scale the magnitude of the normalised templates to the original loading levels.

In this chapter, substation clustering and classification are introduced. 10 distinctive LV substation templates are developed by using the real-time data of 800 monitored substations metered at 10-minute interval over the course of one year (2012-2013). Then, a classification model is designed to assign unmonitored substations to an appropriate template with high statistical confidence. The work is part of a Smart grid demonstration project – LV Network Templates [73] jointly commissioned by WPD

in the UK and the UK's regulator - Ofgem. The developed three-stage LV network load profiling is extensively demonstrated through the LV Network Templates project. This chapter introduces the first two stages and the scaling will be introduced in Chapter 5.

The major contributions of the three-stage load profiling are:

i) it is the first attempt to visualise LV distribution networks by directly clustering and classifying substations rather than aggregating the load profiles of end-users;

ii) Compared to common pattern recognition techniques, the proposed MLR classification does not require any sampled data, which fits well the widely unmonitored LV substations;

iii) The scaling is innovatively designed by cluster-wise regression for network load profiling. By using scaling, the load profiling method can provide not only shape estimation but also more accurate magnitude estimation.

The rest of the chapter is organised as follows: Section 4.3 introduces the LV Network Templates project. Section 4.4 introduces the overall methodology. Section 4.5 describes the clustering techniques and classification model. Section 4.6 discusses the implementation. Section 4.7 analyses clustering results. Conclusions are drawn in Section 4.8.

4.3 LV Network Templates Project

WPD has initiated the LV Network Templates Project in 2012 [73]. The overall aim of the project is to develop a number of common LV substation templates from monitored substations in trial areas, which are applicable to unmonitored LV substations nationwide. They can be used to visualise and understand the conditions of unmonitored LV substations, providing efficient network management with much lower cost.

In order to collect network performance data, WPD deployed monitoring equipment at 800 HV/ LV substations and over 3,500 ends of LV feeders. A selected areas and networks were involved in the project, containing a good mix of geographical

characteristics, customer composition and network topologies. For example, Cardiff is selected as inner city with a larger number of commercial customers and load. Rural areas are represented by regions like Monmouthshire. The geographical area is depicted in Figure 4-1[74].

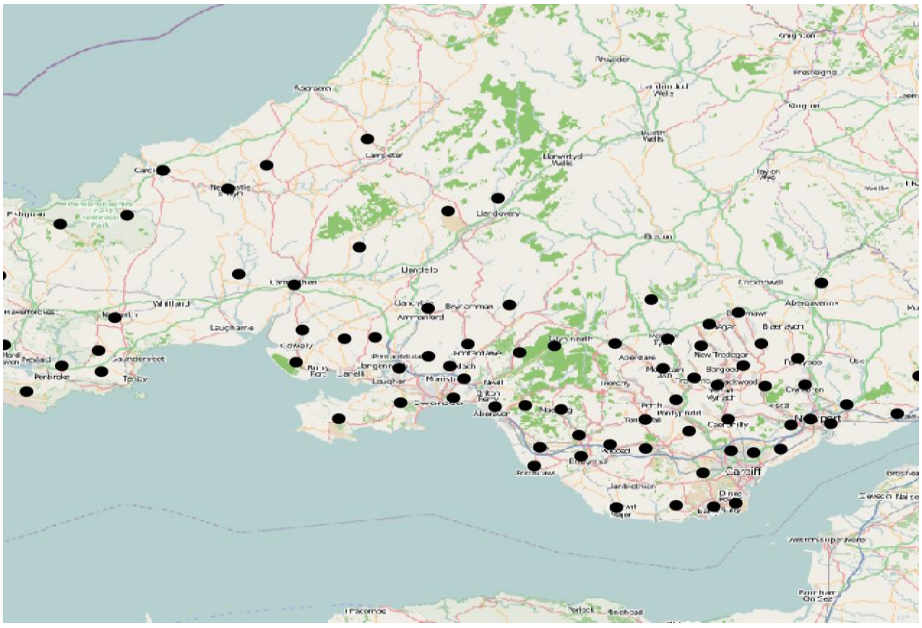

Figure 4-1 Geographical areas of LV Network Template project

Two sets of data are received: fixed data and variable data. Fixed data includes: i) the information of selected LV substation - capacity, connected PV numbers, and the outgoing LV feeder numbers, served customer class and numbers, tariff types and loading levels; ii) LV feeders' information -types, length, and upstream LV substations. The fixed data does not change during the period and it is used to classify LV substations into groups with different characteristics. Variable data includes three-phase voltage, current and real power delivered at HV/LV substations, and three-phase voltage at LV feeder ends, and they are collected at 10-minute interval throughout a whole year.

The first step of the project involves sample size design, data collection and sense checking. 730 LV substations out of 824 have passed sense-checking and been used in this study. Details can be found in Appendix A.

4.4 Overall Flowchart of the Methodology

In this chapter, a combined clustering and classification method is proposed for template development. Firstly, substations are clustered into groups according to their load profile shapes without any pre-knowledge. Based on the clustering results, fixed data is utilised to represent their types and characteristics. MLR model is used to develop a classification tool, which can effectively link the fixed data with cluster memberships of substations. The overall flowchart is summarised in Figure 4-2, which has three major steps.

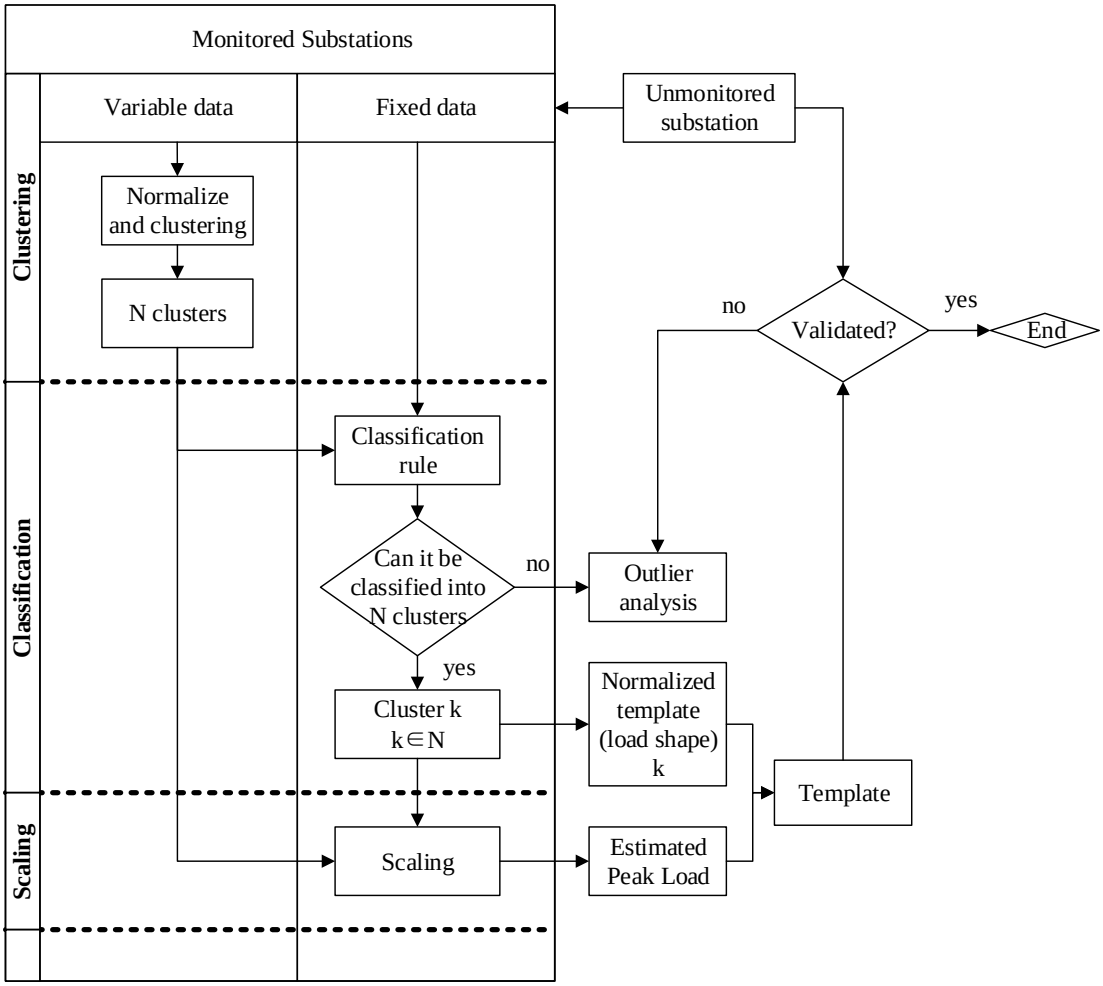

Figure 4-2 Flow chart of the methodology

- Step 1: clustering – Metered load profiles are firstly normalised to represent only shape information. Substations are then grouped according to load profile shapes, producing substation clusters. A normalised template, i.e. averaged load shape

within group, is produced to represent the overall load shape of substations within each cluster (varying over seasons and days);

- Step 2: classification – The relationship between these developed clusters and fixed data of substations is characterised. A classification tool is developed by using the MLR model. It can assign unmonitored substations to each cluster with a certain probability solely based on the fixed data. Thus it is transferable to areas and time periods, for which real-time monitored network condition data is not available;

- Step 3: scaling – Finally, the magnitude of the normalised template needs to be scaled up in order to regain load magnitude information apart from shape information. For un-monitored substations, the scaling process is actually to estimate peak loads by only using fixed data. Scaling will be introduced in Chapter 5.

4.5 Clustering and Classification

This part introduces clustering and classification techniques used for developing LV network templates. Hierarchical clustering is utilised to find the latent groups within metered load profiles of all LV substations. Cluster number is determined by K-means clustering combined with practical considerations. The MLR model is adopted to find classification rules from the fixed data of LV substations.

4.5.1 Hierarchical Clustering for Load Profiles

In cluster analysis, the data can be partitioned into meaningful subgroups, when the number of subgroups and other information about the composition is unknown [75]. With the goal of starting clustering with absolutely no pre-knowledge of the data and ending with a deterministic trace on how each object is clustered into a particular cluster, hierarchical clustering method is chosen for its advantage in achieving these targets.

There are two potential approaches for hierarchical clustering, differentiated by whether building the clusters is performed using a top-down (divisive) or bottom-up (agglomerative) strategy [76]. In the former method, initially all objects are considered to be in a single cluster, and then it is split into smaller clusters iteratively

based on measures of dissimilarity until ultimately each object forms its own cluster. In the latter approach, the initial set-up treats each object being classified as its own cluster, and then different clusters are merged according to similarity measure until ultimately there is again a single cluster. For large dataset, it is often more computationally efficient to use agglomerative hierarchical clustering, where the number of possible merging stage is bounded by the number of observations.

In this study, considering the daily load profiles of a substation (metered every 10 minutes) as a set of 144-element vectors $x = \{x_1, x_2, \ldots x_{144}\}$ which need to be classified, the load matrix X can be written as a N×144matrix in (4-1) for N substations:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,143} & x_{1,144} \\ \vdots & & \vdots & & & \vdots \\ x_{N,1} & x_{N,2} & \cdots & \cdots & & x_{N,144} \end{bmatrix} \tag{4-1}$$

There are three steps in implementing Hierarchical Clustering, which are:

i) The first step- distance calculation

The first step is to obtain the distance matrix D by calculating $d_{k,j}$ the distance between rows x_k and x_j. The most popular choice is the Euclidian distance, which is used in this thesis:

$$d_{k,\,j} = \|x_k - x_j\| = \sqrt{\sum_{i=1}^{144} \left(x_{k,i} - x_{j,i}\right)^2} \tag{4-2}$$

where, x_k and x_j is the daily load of substation k and j.

A dissimilarity matrix is derived to compare the distances between the metered profiles, thus guiding the vectors grouping in the next step. It is a symmetric N×N matrix whose $(i,j)^{th}$ element provides a measure of the dissimilarity between the i^{th} and the j^{th} objects (i,j=1,…,n). The dissimilarity matrix can be written as:

$$D = \begin{bmatrix} 0 & & & & & \\ d_{2,1} & 0 & & & & \\ d_{3,1} & d_{3,2} & 0 & & & \\ \vdots & & & \ddots & & \\ \vdots & & & & \ddots & \\ d_{N,1} & d_{N,2} & \cdots & \cdots & d_{N,143} & 0 \end{bmatrix} \tag{4-3}$$

Dissimilarity d_{ij} reflects the dissimilarity between substations i and j. It should satisfy the following three conditions:

$$\begin{cases} d_{ij} \geq 0 \\ d_{ii} = 0 \qquad i, j = 1,2,...,n \\ d_{ij} = d_{ji} \end{cases} \tag{4-4}$$

ii) The second step-grouping vectors

The second step is to group the vectors x into hierarchical cluster tree by merging together those vectors with the smallest distances. After that a new distance is computed to all the other vectors or clusters. The process of forming new clusters is repeated until only one-cluster remains. In a set of N vectors N-1 merging operations are needed.

In this study the Ward distance is used for calculating the distance between clusters. For two clusters A and B, their distances are defined as (4-5)

$$d_{A,B} = \frac{1}{|A||B|} \sum_{x_a \in A} \sum_{x_b \in B} \sqrt{\sum_{i=1}^{144} (x_{a,i} - x_{b,i})^2} \tag{4-5}$$

where, x_a is the daily load of substation α from cluster A and similar x_b represents daily load of substation b from cluster B.

Starting with M clusters, each cluster would contain a series of data from the monitored substations. When the dissimilarity measure has been calculated between all clusters, two clusters that are most similar (with the least distance) are merged, leaving M-1 clusters. This can be repeated until there is only one cluster left and the result is a binary tree of 2M-1 clusters. The above-described process forms the Hierarchical Cluster tree, which is stretched as a dendrogram as in Figure 4-3.

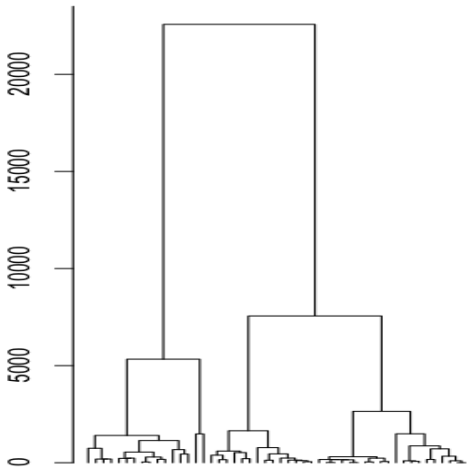

Figure 4-3 Example of a dendogram

iii) The third step-division of clusters

The last step of the clustering is to divide the cluster tree into coherent cluster groups. It can be achieved by cutting the hierarchical tree at an appropriate point, which can be determined by the number of clusters needed to be developed or the distance between clusters.

4.5.2 K-Means Clustering for Cluster Number Determination

It is essential to determine the optimum cluster number in clustering. As LV networks cannot be pre-classified into macro-categories (e.g. residential, commercial or industry), the clustering needs to handle a large volume of metered network data through the study year. K-means clustering is chosen here because it can accelerate the repeated assessments of different numbers of clusters [77]. It can provide a quick assessment of the range of possible optimum number of clusters. For daily load profiles of N substations described in (4-1), it proceeds by selecting *k* initial cluster centers $(C_1^{(1)}...C_k^{(1)})$ and then iteratively refines them through:

i) Each instance x_i is assigned to its closest cluster center to form *k* data set;

$$S_j^{(1)} = \{x_i : \left\| x_i - c_j^{(1)} \right\|^2 \le \left\| x_i - c_n^{(1)} \right\|^2 \forall 1 \le n \le k\} \quad (4\text{-}6)$$

ii) Each cluster center C_j is updated to be the mean of its constituent instances. The algorithm converges when there is no further change in assignment of instances to clusters.

$$c_j^{(t+1)} = \frac{1}{\left|S_j^{(t)}\right|} \sum_{xi \in S_j^{(t)}} x_i \tag{4-7}$$

Different cluster numbers are assessed in terms of the dissimilarity within each group. The dissimilarity within cluster is represented by the Sum of Squares of the Errors (SSE) between LV substation load profiles within groups. For example, when cluster number is *k*, the objective set is $S = \{S_1, S_2, \ldots S_k\}$ with the centre $C = \{C_1, C_2, \ldots C_k\}$. The total internal distance of all clusters is calculated in (4-8) as the SSE

$$Dis = \sum_{j=1}^{k} \sum_{x_i \in s_j} (x_i - C_j)^2 \tag{4-8}$$

4.5.3 Multinomial Logistic Regression for Classification

After creating a set of clusters, a classification tool is required to assign unmonitored LV substations to the developed cluster groups. It intends to assign them to the most similar clusters by only using their available fixed data.

Logistic Regression is a regression model designed to predict the outcome of a categorical dependent variable [78]. It is used here to design classification tools. Given a set of independent variables (fixed data) *z* and the dependent variable (substation cluster) *y*, the regression coefficients are b_0 and b_1. The model is trained to produce probability of binary outcomes by logistic function

$$P(y=1) = F(z) = \frac{1}{1+e^{-(b_0+b_1 \times z)}} \tag{4-9}$$

The result in (4-9) is between 0 and 1, which is interpreted as the probability of dependent variable *y* belonging to 1. The probability for *y* belonging to 0 is $P(y=0) = 1 - P(y=1)$

For more than two categories (clusters), MLR is used to analyse the relationship between cluster membership and fixed data. For *K* clusters, the model composes of *K*-

1 logistic regression equations, all of which use cluster *K* as the baseline. The probability of substation Y_i belonging to cluster *n* can be determined by (4-10) and (4-11). A substation is allocated to the cluster with which it has the highest probability.

$$P(Y_i = n) = \frac{e^{b_n \times z_i}}{1 + \sum_{k=1}^{K-1} e^{b_k \times z_i}}, \qquad n = 1,2,...K-1 \tag{4-10}$$

$$P(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{b_k \times z_i}} \tag{4-11}$$

where, $z_i = (z_{i,1}...z_{i,m})$ is the i^{th} set of *m* independent variables (fixed data), $b_k = (b_{0,k}...b_{m,k})$ is the regression coefficient for cluster *k*; and $P(Y_i = n)$ is the probability that the i^{th} substation belongs to cluster *n*. The parameters are estimated by maximum likelihood Estimation (MLE). Detailed derivation and algorithm can be found in Appendix B.

4.6 Implementation

This section details load profile normalization, the determination of cluster number, and the development of classification tool.

4.6.1 Normalization

The units of the data used as input for the clustering to a large extent determine the nature of the clusters. As transformer ratings in LV substations vary substantially, the direct use of the measurements in active power (in kW) will produce clusters that only reflect the magnitudes of load but not their shapes. Therefore, normalization of the data is necessary in order that the developed clusters can reflect load patterns/variations within a day. Along with directly using measurements, two other normalization approaches are used:

- Magnitude – the data on the original measurement scale is used;
- Normalised I – the data is normalised according to the maximum value of real

power delivered at each substation over the entire period of study (annual peak);

- Normalised II – the data is normalised according to the maximum value of real power delivered at each substation of each day (daily peak).

All three types of inputs are tested for clustering and some important findings are observed. When the first approach is used, the developed clusters are dominated by the magnitude of load associated with each substation and therefore the profile patterns within days are not properly reflected. The second approach can to some extent overcome the disadvantage of the first approach, but it is affected by daily or seasonal changes in the maximum values, thus producing a set of clusters that largely reflect the overall effect of time.

In the UK, winter demand usually dominates annual energy consumption due to electrical heating and lighting. Figure 4-4 shows the annual load of a metered substation normalised by the second method. It clearly shows that spring, summer and autumn have similar peaks but winter has a sharp peak. The difference of load magnitudes between seasons will dramatically affect the clustering accuracy. In this case, clustering of substations will be dominated by winter load

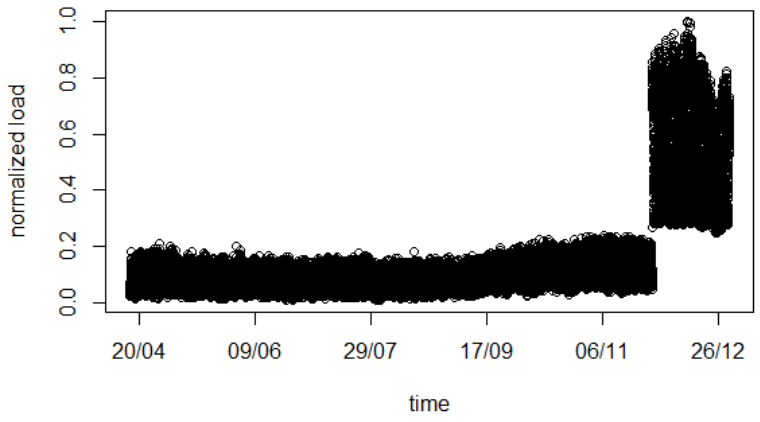

Figure 4-4 Annual load of a selected substation by normalised I method

The third approach is specifically designed in developing clusters in order to detect the patterns of demand within a calendar day. The clusters are constructed based on the difference in demand over a 24-hour period, which are further refined by separating weekdays and weekends (Saturdays and Sundays treated separately) and seasons (Spring, Summer, High Summer, Autumn and Winter). Further, anomalous

days are more likely to have critical peaks/troughs as they can dominate/distort the whole year load profiles. The normalization II is on daily basis and therefore anomalous days will not influence other normal days. All days are clustered together and thus anomalous days with very different load shapes will be automatically clustered to other groups, or labelled as outliers. The normalised load vector of substation *i* in a single day is

$$Norm\{P_{ij}\} = \frac{P_{ij}}{\max\{P_{i1},...,P_{i144}\}} \mid j = 1,...,144 \tag{4-12}$$

where, P_{ij} is the metered load of substation *i* at time *j* of a day, and $Norm\{P_{ij}\}$ is the normalised load.

4.6.2 Determining Cluster Number

Theoretically, there could be one cluster to represent all LV substations, but the precision would be extremely poor. At another extreme, there could be as many clusters as the number of all LV substations, i.e. each cluster represents only one LV substation. In this case, the precision is high but it is impractical due to high management burden. Here, the optimum cluster number is determined by considering a combination of clustering and classification accuracy, including:

- The comparison of the variations 'within' versus 'between' clusters is achieved by K-means clustering as introduced in Section III. Figure 4-5 shows the Sum of Square Error (SSE) within groups against cluster number, showing decreasing benefits in terms of SSE drop obtained by increasing cluster number.

- The classification mainly is the predictive accuracy of allocating unmonitored substations to clusters using the classification tool. A substation is allocated to the cluster with the highest probability obtained by MLR. The accuracy is assessed by comparing predicted cluster membership with actual membership in clustering.

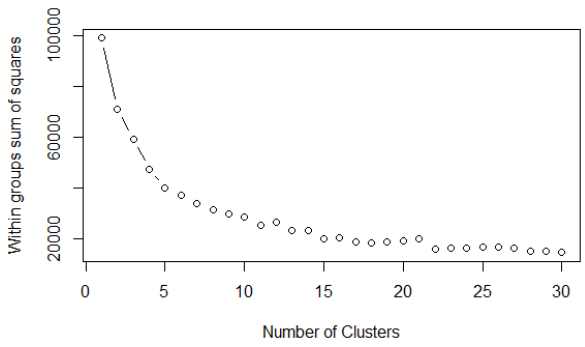

Figure 4-5 Within-group sum of squares errors against number of clusters

4.6.3 Classification Tool

A classification tool is needed to link the developed templates with fixed data so that it can predict the cluster membership of unmonitored substations by only using fixed data. The selection of factors for the MLR model is made by using a combination of statistical significance - improved model fitness (likelihood) and the ability to accurately predict cluster membership. The model fitness can be improved by adding more parameters. However, it may cause over-fitting if too many low-related parameters are included, leading to inaccurate prediction performance. The trade-off is made by Bayesian Information Criterion (BIC), which introduces a penalty term with the number of parameters, shown in (4-13). The model with the lowest BIC is chosen.

$$BIC = -2 \cdot \ln L + k \cdot \ln(n) \tag{4-13}$$

where, L is the maximum likelihood value which can be obtained from the assessed model; k is the number of parameters and n is the sample size.

The fixed data used as inputs for classification tool development is as follows:

- Number of customers in each class
- Estimated annual consumption
- Transformer type
- Percentage half hourly load

- Total feeder length
- Transformer rating
- Number of feeders
- Percentage of overhead lines/underground cables

The output of classification consists of a set of probabilities, each of which indicates the likelihood that a substation belongs to a particular cluster. The predictive accuracy of classification tool is calculated by comparing cluster memberships between clustering and classification predictions. The tested LV substations are mapped to clusters through the classification tool by only using their fixed data. The accuracy is calculated as the proportion of LV substations which can be mapped by the classification tool to the same cluster as the clustering does.

4.7 Demonstration and Results

The demonstration of the designed methodologies is conducted on the smart grid demonstration project – LV Network Templates commissioned by WPD and Ofgem.

4.7.1 Number of clusters

In clustering test, the "knee point" of SSE shown in Figure 4-5 indicates that the optimal cluster number should be between 5 and 15. Considering that bigger cluster number produces higher accuracy, the number is reduced to 7-15.

In classification, using a single cluster will result in 100% accuracy, as every substation can be correctly allocated to the single cluster. The cluster number from 2-20 is investigated in this study and the accuracy index introduced in Section V is calculated. Table 4-1 shows the predictive accuracy associated with different cluster numbers. It shows an increase accuracy from 75.5% when 7 clusters are developed to a maximum of 82.2% when 10 clusters are used. However, there is a 'turning-point', where the increasing number decreases the predictive accuracy, with a sharp decrease for 11 clusters (70.6%). The results in Table 4-1 indicate the optimal cluster number should be 10, which is actually used in the study.

Table 4-1 Predictive accuracy of classification with various cluster number

Number of clusters	Accuracy (%)
7	75.7
8	75.5
9	76.9
10	**82.2**
11	70.6
12	70.5
15	65.3

4.7.2 Clusters and Normalised Templates

After determining the cluster number, which is 10, agglomerative hierarchical clustering is used to cluster the training set of substation load profiles. After clustering, it is found that substations within each cluster have very similar characteristics in terms of customer mix and fixed data. A summary of fixed data of substations within each cluster is provided in Table 4-2. The percentage indicates the annual energy consumption of certain type of customers. The third column illustrates the number of substations used for clustering in each cluster, based on high summer weekday load profiles.

Figure 4-6 gives a more detailed description of cluster 1 template (black line) and the individual load profiles of substations within cluster 1 that generate it. The average normalised load pattern of substations within each cluster, season and day type is defined as template. Clearly, template 1 to much extent can represent the variations within those load profiles, with daytime having relatively flat peak but night-time having trough demand.

A calendar year is divided into 5 seasons (spring, summer, high summer, autumn and winter) and one week is divided into 3 typical days (weekday, Saturday and Sunday). It is consistent with the practice of the power industry and settlement market in the UK. For each day type in a season, 10 templates are divided, as listed in Table 4-2. Thus, there are 150 templates in total (5 seasons × 3day types × 10 templates = 150 templates), representing a large variety of LV substations.

Table 4-2 Ten LV network templates

Cluster	Description of fixed data	Substation number (High Summer Weekday)
1	High I&C Dominance	62
2	Modest Domestic Dominance (~60%) (Suburban)	37
3	Modest Domestic Dominance (~60%) (Urban)	255
4	High Domestic Dominance (~90%) (Modest Customer Size ~170)	172
5	High Domestic Dominance (~90%) (Low Customer Size ~70)	66
6	Very High I&C Dominance (~90%)	44
7	Modest Domestic Dominance (~60%) (Rural)	41
8	Industrial Flat	27
9	Domestic Economy 7 Dominance (~65%)	21
10	Lighting	5

The mix of substations in different typical days and seasons are handled in the following ways: i) for a day type in a season (e.g. winter weekday), load profiles of different substations and different days are clustered together. Substations with majority of days (over 80% of studied days) clustered into the same cluster are defined as typical substations and used in the templates. By contrast, substations with days evenly separated into different clusters are defined as outliers, thus not used in the templates. Detailed outlier analysis in terms of both load shape and scaling can be

found in Appendix D; ii) as the clustering is independent, a substation may be clustered into different clusters in different seasons and days. Therefore, the coefficients in the classification (also in scaling) step are all specified for each season and day type.

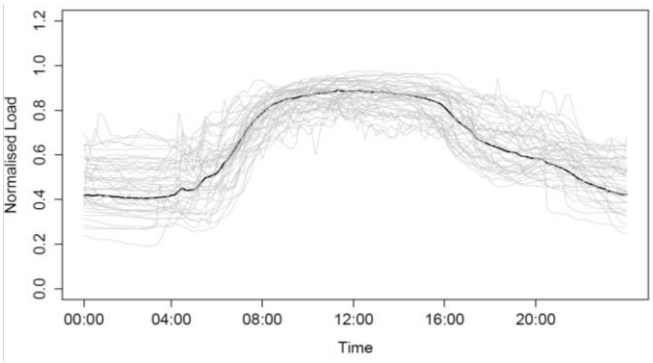

Figure 4-6 LV substation template and within-cluster variability

Due to limited space, this chapter only provides the developed templates representing load profiles in weekdays of high summer. The rest of the scenarios can be found in Appendix C. In Figures 4-7 to 4-11, the black lines are the normalised daily load templates for each cluster and the red dashed lines are the standard deviation of all substations' load profiles within cluster. A standard deviation indicates the variations of LV substation load within the same cluster at different time of day. It is noted that most standard deviations are below 0.2 (relative to 1.0), showing small variations within profiles. The developed templates are able to represent load shapes of LV substations that have similar fixed data without compromising much precision.

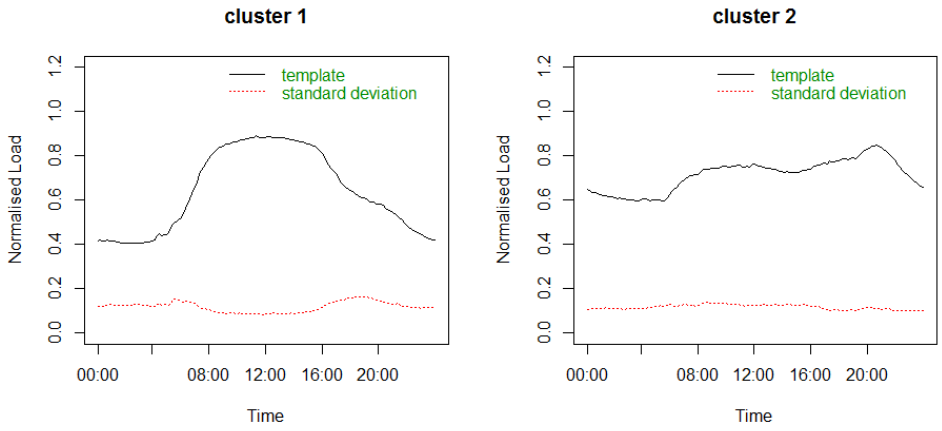

Figure 4-7 Templates and standard deviations for clusters 1 and 2

Cluster 1 in Figure 4-7 is largely commercial dominated substations with a relatively high flat demand during daytime and lower demand overnight. The major composition of LV substations in this cluster is commercial customers, and their average peak is around 0.8 unit. Cluster 2 predominately comprises of substations dominated by those serving domestic customers and perhaps a certain proportion of commercial customers. The profile is relatively flat during daytime, but peaks at around 19:30pm. The analysis of the fixed data finds that most of the LV substations in cluster 2 are located in suburban areas.

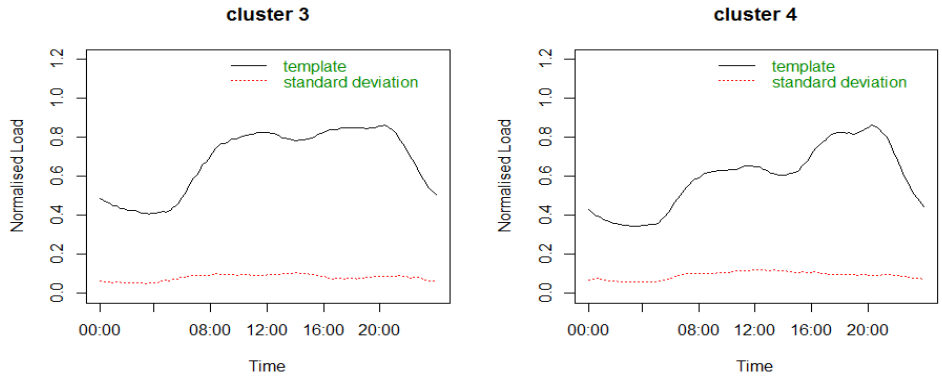

Figure 4-8 Templates and standard deviations for clusters 3 and 4

Clusters 3 and 4 are depicted in Figure 4-8. Most LV substations in cluster 3 are located in urban areas, serving a mix of domestic and commercial customers. The shape of cluster 3 in daytime is similar to that of cluster 1, but has an increasing demand during night time due to a higher proportion of domestic customers. Its peak occurs at around 20:30pm with the value of around 0.9 unit. Cluster 4 comprises largely of domestically dominated substations. Compared to clusters 2 and 3, the profile of cluster 4 is relatively flat and low during daytime.

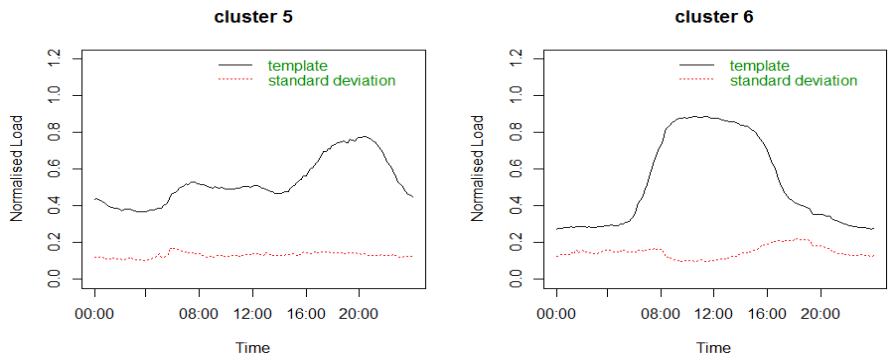

Figure 4-9 Templates and standard deviations for clusters 5 and 6

Cluster 5 contains highly domestic dominated substations but with a small number of customer around 70. The shape is similar to that of cluster 4, but the daytime load is much lower and more flatten. The evening peak is also lower at about 0.8 unit. Cluster 6 represents load profiles of commercially dominated substations, whose pattern is similar to that of cluster 1. The proportion of commercial customers in cluster 6 is much higher, which is reflected by higher peak and lower night load.

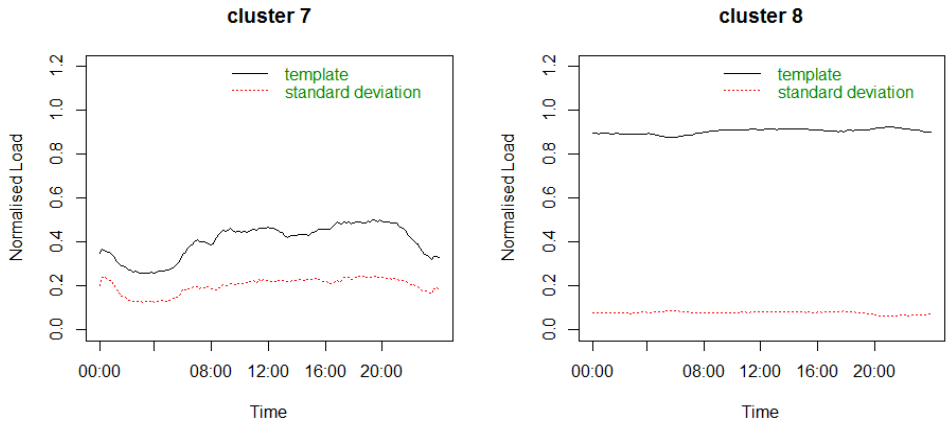

Figure 4-10 Templates and standard deviations for clusters 7 and 8

Cluster 7 largely contains a mix of domestic substations and small commercial substations in rural areas with low demand. There are two slight peaks, the first of which appears around 12:00 pm driven by commercial demand and the other occurs at approximately 20:00 pm triggered by domestic demand. Cluster 8 comprises of substations heavily dominated by industrial customers, whose load shape shows consistently high loading through at day.

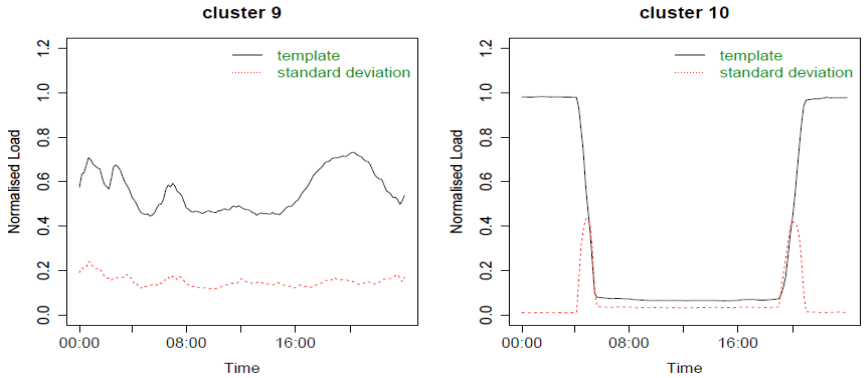

Figure 4-11 Templates and standard deviations for clusters 9 and 10

Cluster 9 represents load profiles of substations with significant proportion of Economy 7 customers, who have a lower electricity rate during night and higher rate in daytime. There is a night peak at around 1:00 am and this night peak becomes as high as nearly 1.0 unit in winter time, when Economy 7 customers tend to use electricity heating. Cluster 10 exclusively comprises of substations serving motorway communication/ lighting pillars.

4.7.3 Classification Tool

In order to make the templates easier to use by other DNOs to understand the conditions of their unmonitored LV substations, an interface tool is developed to match LV substations to the most likely clusters only based on fixed data. By inputting its fixed characteristics including customer mix and network structures, etc. the probability of it belonging to each cluster is calculated by MLR. Initially, benchmark template of the highest probability cluster is given. The tool is tested by all substations in this project. The fixed data of all trial substations are used to predict cluster membership. With 10 clusters, 82.2% substations are predicted with the highest probability clusters, being the same as those derived from clustering.

4.8 Chapter Summary

This chapter proposes a three-stage network load profiling method to provide a direct but economical way to visualise LV networks. It consists of three major steps: clustering, classification and scaling, where the first two are introduced in this chapter. LV network templates are developed by using the metered real-time data from selective areas that are representative. By demonstrating on a practical trial UK smart grid project – LV Network Templates, 10 network load profile clusters with different load shapes are produced. A classification tool is developed to assign un-monitored substations to the appropriate clusters by only using fixed data.

The templates developed in this chapter only contain load shape information, but the magnitudes of load profiles need to be known. The load profile magnitude identification (scaling), validation, and discussion on the application of the developed LV templates will be covered in Chapter 5.

Chapter 5

Time-series Load Profiling for LV Networks: Peak Load Estimation

THIS chapter proposes an innovative method to improve the accuracy of peak estimation for LV networks. It aims to scale the templates to the real loading levels, reflecting both load shape and magnitude.

5.1 Introduction

Chapter 4 has developed classification rules can allocate unmonitored substations into the most appropriate clusters and thus provide indicative templates. These normalised load templates only indicate the shape of substation electricity usage, but not actual loading level. In practice, the loading levels across LV substations vary to a great extent [7], making it impossible to study a large number of LV substations on the same scale, e.g. unified scale [0, 1]. Scaling is therefore needed to adjust the normalised templates back to the natural magnitude for each substation. Scaling the normalised templates allows a good range of loading levels to be covered by the templates, thus adding great flexibility to the use of the templates.

This chapter proposes a novel contribution factor approach to predict diversified daily peak load of LV substations. The contribution factor for each LV template developed in Chapter 4 is determined by a novel method - Clusterwise Weighted Constrained Regression (CWCR). It takes into account the contribution from different customer classes to substation peaks, respecting the natural difference in time and magnitude between LV substation peaks and the variance within the templates. In CWCR, intercept and coefficients are constrained to ensure that the resultant coefficients do not lead to reverse load flow and can respect zero-load substations. Cross validation is developed to validate the stability of the proposed method and prevent over fitting. The proposed method shows significant improvement in the accuracy of peak estimation over the status quo across 800 substations of different mixes of domestic, industrial and commercial (I&C) customers.

5.2 Problem and Proposed Solution Statement

The templates in Chapter 4 are developed from the normalised load profiles where the original load profiles are normalised by their daily peaks. Thus scaling is used to estimate and recover substation daily peaks. The challenge in doing this is that the peak estimation has to solely rely on readily available fixed data so that it is applicable to unmonitored substations. In practice, the power industry typically adopts P-Q method [15] to estimate the annual peak of LV substations. This method only provides an annual estimation and its accuracy is highly compromised by using many statistical assumptions, such as a certain level of standard deviation added onto the

mean demand [15, 55]. In literature, other methods [37, 79] have been reported to estimate system or HV substation peak load. They largely depend on the sufficiency of historical load data, which is usually unavailable at LV substations.

A well-known peak estimation method for LV substations is the kWh-to-peak-kW Conversion factors and Diversity factors (C-D) method [80]. It takes advantage of the information of customers' bills to aggregate individual customer's peak to substation peak. Based on this concept, recent work adopts advanced techniques including fuzzy regression [81], fuzzy inference [82] and artificial neural network (ANN) [58] to handle uncertainties, narrow confidence intervals and improve the accuracy. However, all these methods do not consider the diversifications in customer composition and structures of LV substations. For different types of LV substations, customer's peak has various coincidence degrees, in relation to substation peak and it is therefore fairly inaccurate to use one single model to estimate the peaks for all LV substations.

This chapter proposes a Clusterwise Weighted Constrained Regression (CWCR) approach for LV substation peak demand estimation based on fixed substation information. It innovatively develops cluster-specified estimation model, which for the first time addresses the variances of customer peak load's contributions to the peaks of different types of LV substations. Based on substation clustering and classification, peak estimation parameters are derived for each cluster by the clusterwise regression model. The method is weighted by variances within clusters and constrained with zero intercept and non-negative coefficients, considering practical constraints and statistical accuracy. Cross validation is then used to prevent over-fitting and validate the applicability of the proposed method. The predicted results are compared with those from the traditional industry P-Q, kWh-to-peak-kW conversion method and diversity factors (C-D) method. The comparison shows that the proposed method can achieve substantial improvement in accuracy in estimating peak demand for LV substations over other methods.

The rest of the chapter is organised as follows: Section 5.3 introduces the concept of contribution factor. The CWCR approach is proposed in Section 5.4 and its implementation is discussed in Section 5.5. In Section 5.6, the proposed scaling

methods are demonstrated and cross validated. The use of scaled templates is discussed in Section 5.7. Conclusions are drawn in 5.8.

5.3 Rationale of Contribution Factor

This section introduces the diversity of LV substation as the latency classes underlying the datasets. A contribution factor is developed to mathematically describe the question and simulated examples are used to illustrate it.

5.3.1 Latency in LV Substation Class

The issue of latent LV substation classes in peak estimation has hardly been studied in previous research due to its complexity. One LV substation usually serves different type of customers, but not every customer's peak load coincides with the aggregated substation peak. Customers contribute differently to LV substation peak. Moreover, even the same class of customers can contribute differently due to the diversification of LV substations. Taking the results in Chapter 4 as an example, Figure 5-1 plots the daily load profiles of a typical domestic customer, a sample domestic dominated cluster (cluster 5) and a sample I&C dominated cluster (cluster 6). As seen, the domestic customer's peak coincides with that of cluster 5 but not cluster 6. As a result, a domestic customer contributes more to the peak of its substation if it is a domestic dominated substation, but less to the substation dominated by I&C customers.

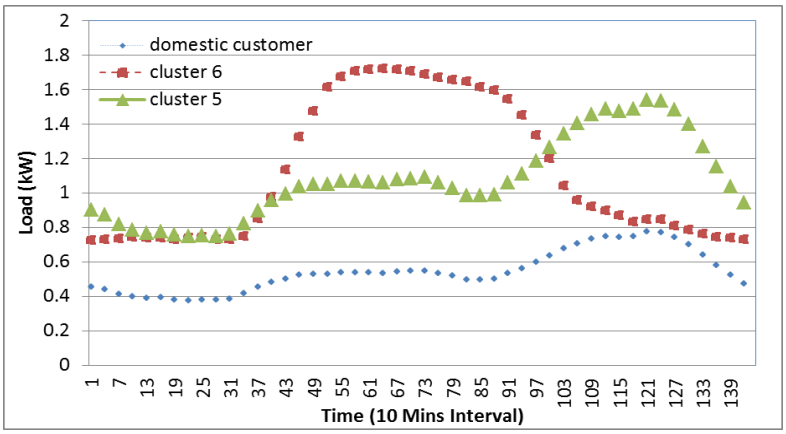

Figure 5-1. Load profiles of a typical domestic customer, cluster 5 and cluster 6

This latency should be reflected in LV substation peak estimation. Otherwise, it is inappropriate to use one single set of factors or models with deterministic parameters

to estimate peaks for all LV substations. The peaks of various types of LV substations should be appropriately partitioned and separately estimated by different models and parameters.

5.3.2 Contribution Factor

This chapter for the first time proposes a contribution factor to address the contribution from a particular customer to the peak of different type of LV substations. A contribution factor is designed to describe the coincidence level between different type of customers and substations. It is defined as the ratio between aggregated peak of one class customers and their contribution to different types of substations in (5-1)

$$\begin{cases} F_{ik} = \dfrac{S_i(t_q)}{S_i(t_p)} \\ S_i(t_p) = \max(S_i);\ L_k(t_q) = \max(L_k) \end{cases} \tag{5-1}$$

where, F_{ik} is the contribution from class i customers to the peak of cluster k; t_p is the time when the aggregated peak occurs and t_q is the time when substation metered peak occurs; $S_i = [S_i(t_1), S_i(t_2), \cdots, S_i(t_n)]$ is the aggregated daily load of class i customers; the loading level of cluster k substations is $L_k = [L_k(t_1), L_k(t_2), \cdots, L_k(t_n)]$; $t = [t_1, t_2, \cdots, t_n]$ is the time intervals of daily load profiles; .

The contribution factor can provide cluster-specified peak estimation for LV substations. Based on the C-D method in (2-3), the peak load can be estimated by (5-2).

$$\max(L_k) = \sum_{i=1}^{I} E_i \cdot \frac{C_i}{D_{in}} \cdot F_{ik} \tag{5-2}$$

5.3.3 Illustration of Contribution Factor

Two groups of 100 domestic dominated substations and 100 I&C dominated substations are taken as examples to illustrate the impact of contribution factor. Each point in Figure 5-2 represents one LV substation. The X axis represents the

aggregated peak $S_i(t_p)$ of all domestic customers and Y axis is the actual LV substation peak- $S_i(t_q)$. The slope of the solid red lines is contribution factor defined by (5-1). Clearly, two groups of LV substations exist: for domestic dominated substations, domestic customers contribute nearly 100% to their peak, which forms the upper group with a slope close to 1. By contrast, the lower group represents I&C LV substations. Less than 50% of domestic customers' peak actually contributes to these substations' peak. If the peaks of these 200 LV substations are estimated by the traditional C-D method, shown by the dash line in the centre, the accuracy is compromised.

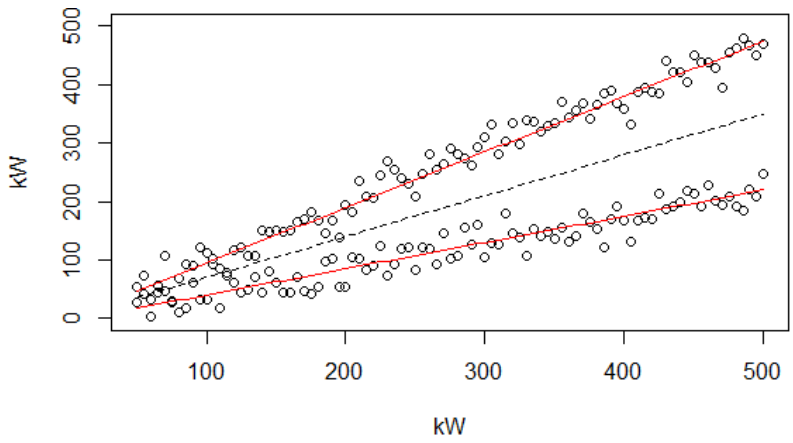

Figure 5-2. Individual customer contribution to substation peak

Figure 5-2 illustrates that it would be very inaccurate to estimate substation peak by using a single regression model. By contrast, the data should be partitioned by underlying classes and estimated separately. In this way, the 200 substations can be categorised into two categories, as highlighted by the two solid red lines. Thus, it is more appropriate and accurate to treat the two groups separately, justifying the need to develop two regression models for them.

5.4 CWCR Based Scaling Method

This section proposes a new method to estimate LV substation peak, which can identify the contribution from different customers and fundamentally increase the estimation accuracy. The contribution factor is included in the approach, which is obtained by CWCR.

Based on the 10 clusters developed in Chapter 4, substations with similar load shapes are clustered and their peak loads are expected to appear at similar time as well. Regression models can be developed for each cluster to interpret the relationship between fixed data and peak load. The basic idea is to transform the normalised templates developed in Chapter 4 into scaled templates for each type of LV substations. This process is implemented solely by using the available fixed data so that the developed templates and peak estimation methods can be widely applied to un-monitored LV substations, only whose fixed data is normally available. The proposed methodology consists of the following four steps.

i) Step 1: Substations are split into 2 sets: training set and testing set. Training set is used to develop the model and it is then assessed by the testing set.

ii) Step 2: Clustering is conducted by different seasons and day types, and the scaling follows the same routine. Starting with the LV substations in the training set, the relationship between substation peak and fixed data is developed by CWCR. Within LV substation templates developed in Chapter 4, weighted and constrained regression is combined into clusterwise regression, considering practical and statistical conditions. The detailed development process will be discussed in Section V.

iii) Step 3: The proposed method is validated by the substations in the testing set. The estimated peaks are compared with real metered values, where the accuracy is quantified in terms of residuals and R-square errors.

iv) Step 4: The whole data set is then re-split into new training and testing sets. The whole process is repeated until every substation is used in both training set and testing set. Cross validation is discussed in Sections V and VI.

5.5 Mathematical Formulation

Based on the proposed methodology, the average daily peaks of LV substations are estimated solely by using their fixed data, which are asessed with the data of 800 monitored LV substations provided in Chapter 4. This section explains the implementation of CWCR and cross validation. The peaks from the designed method

are also compared with those estimated by the C-D and P-Q methods introduced in section II.

5.5.1 Clusterwise Weighted Constrained Regression

In order to properly separate data and conduct estimation by different models, clusterwise regression has widely been used [83, 84]. Constrained regression and weighted regression [85] have also been developed to adjust practical situations. This chapter proposes CWCR by combining clusterwise, constrained and weighted regression. For J metered substations in cluster k, with a total of I customer classes, the peak can be estimated by (5-2), where $max(L_k)$ and E_i are known. C factors, D factors and F factors are to be determined. The estimation of coefficients are written in (5-3)

$$L_{jk} = \sum_{i=1}^{I} E_{ji} \cdot b_{ik} + e_{jk} \tag{5-3}$$

where, j is substation index, i is the length of each independent variables, L_{jk} is the peak load for substation j from cluster k, E_{ji} is the annual consumption of the i^{th} class of customer served by substation j, cluster specified scaling coefficients b_{ik} represents $\frac{C_{ik}}{D_{ink}} \cdot F_{ik}$ in (5-2), n is the number of customers in class i and, e_{jk} is the error for substation j.

Usually, the coefficients b_{ik} can be estimated by Ordinary Least Squares (OLSs), but practical and statistical constraints need to be considered for CWCR.

i) The intercept should be zero as there would be no load if customer number is zero;

ii) Although customers with distributed generation may produce inverse power flow, they are not included in the total customer classes in (5-3). All customer classes are assumed to contribute positive load to the substation load and therefore the coefficient b_{ik} in (5-3) should be non-negative;

iii) Within each cluster, there are variances between the normalised templates and

substations' load shapes. In order to design the peak estimation model for the normalised templates, a weight w_{jk} to reflect the variances within clusters in (5-4) is assigned to substation j in cluster k;

$$w_{jk} = \frac{1}{\sigma_{jk}^2} \tag{5-4}$$

where, σ_{jk} is the variance between substation j to normalised template k.

The regression model in (5-3) is then converted into a weighted non-negative least square problem, which can be solved as an optimization problem described in (5-5) and (5-6).

$$\arg\min_{\bar{b}_k} \; f_k(\overline{b_k}) = \sum_{j=1}^{J} w_{jk} \cdot e_{jk}^2 \tag{5-5}$$

$$s.t.\ \overline{b_k} > 0 \tag{5-6}$$

where, $b_k = [b_{1k}, b_{2k}, \cdots, b_{ik}]$ denotes the vector of cluster specified scaling coefficients b_{ik} in (5-3):

The objective function (5-5) can be further expanded into (5-7). The problem can be solved iteratively by using analytical approaches [86] and it has been shown that the iteration always converges.

$$f_k(\overline{b_k}) = \sum_{j=1}^{J} w_{jk} \cdot \left| L_{jk} - \sum_{i=1}^{I} E_{ji} \cdot b_{ik} \right|^2 = \left\| (\overline{L_k} - \overline{E}\,\overline{b_k}) / \overline{\sigma_k} \right\|^2 \tag{5-7}$$

5.5.2 Cross Validation

Over-fitting is a term referring to that the regression model requires more information than the data can actually provide in order to represent the true relationship in observations [87]. It can occur when a model is initially fit with the same data as was used to assess the fitness. Thus, it is essential to validate the applicability of the designed method to other LV substations.

Due to lack of data, cross validation is adopted here to assess the applicability of different models. It tends not to use the same set of LV substations in both building and validating. This cross validation randomly takes some LV substations as testing set before building the regression model. The substation set left, called training set, is used for regression model development. The coefficients derived from the training set are validated by the testing set. Cross validation repeats this process to assess the over-fitting problem.

N-fold cross validation is adopted here, which randomly splits the whole substation set into N folds. Every time, one fold is taken out as testing set while the rest N-1 folds are used as training set. The process repeats N times until every fold has been used as testing set. If the coefficients from the CWCR model are fit for all testing sets, they should be statistically applicable to other substations with similar fixed data.

5.5.3 Tests of P-Q Method and C-D Method

In the UK, the industry traditionally uses annual energy consumption to estimate peak load for network design [15, 55]. The basic idea is by using statistic methods to convert annual consumption of LV substations to winter mean peak demand. The targeted peak is calculated by adding a certain level of standard deviation onto the mean peak. C-D method [56] adopts kWh-to-peak-kW Conversion factors (C factor) and Diversity factors (D factor) to estimate LV substation peak based on customer billing cycle kWh consumption. The detailed reviews of these two methods were introduced in Chapter 2.

The coefficients of the P-Q method can be derived from national public information like Common Distribution Charging Methodology (CDCM) [11]. It to a large extent represents the average estimation at national level. The variances are taken from typical values [15]. In (2-2), two standard deviations (β=2) are taken because usually peak load is slightly higher than two standard deviations above the mean demand. When designing LV networks, DNOs usually allow a probability level of around 10% overloading, exceeding the design capacity due to economic reasons. It equates to 1.28 standard deviation above the mean demand (β =1.28).

As for C-D method, the coefficients are derived by Ordinary Least Square (OLS)

regression by using the same training set for CWCR so that a more objective comparison can be achieved. The results of both CWCR and C-D methods will be compared by the same testing sets in the following section.

5.6 Results

5.6.1 Fitness Comparison

The peaks of LV substations are estimated only by using their fixed data based on the proposed CWCR, which are also calculated by using P-Q and C-D methods introduced in Sections II and IV. Results obtained by these methods are compared with each other and with the real metered peaks in terms of goodness of fit.

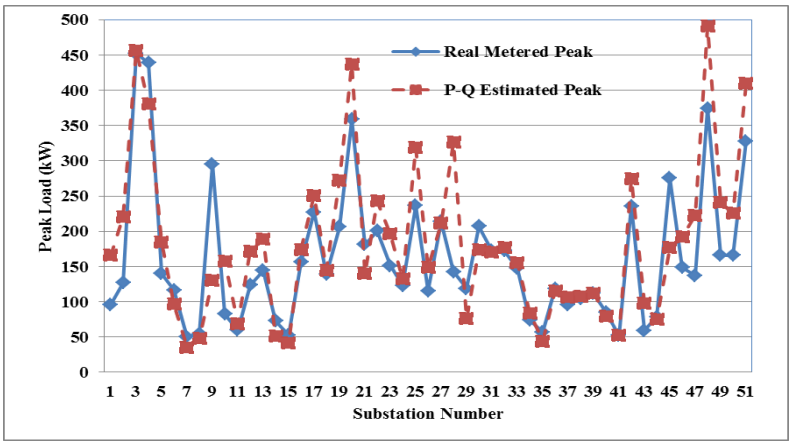

Figure 5-3 Ratio of metered and estimated peak by industry P-Q method

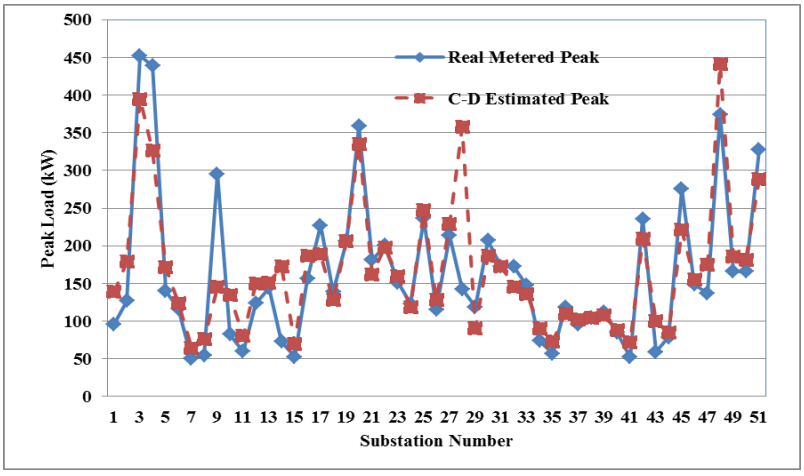

Figure 5-4 Metered and estimated peaks by C-D method (coefficients from OLS regression)

The comparisons are implemented on all trial LV substations. Here only 50 are randomly selected for illustration purpose due to limited space. The estimated peaks from P-Q method and metered peaks are plotted in Figure 5-3. Each point along X axis represents a LV substation, and its estimated and real peaks are depicted by a red dash square and a blue solid point respectively. Figures 5-4 and 5-5 compare the estimation of peak demand for LV substations by OLS-based C-D method and CWCR. As seen, the estimations from P-Q and C-D methods generally follow the real peaks, but the errors can be still seen and sometimes they are rather obvious. In Figure 5-5, the CWCR estimations almost overlap with the real peaks, which clearly show the improvement of in accuracy.

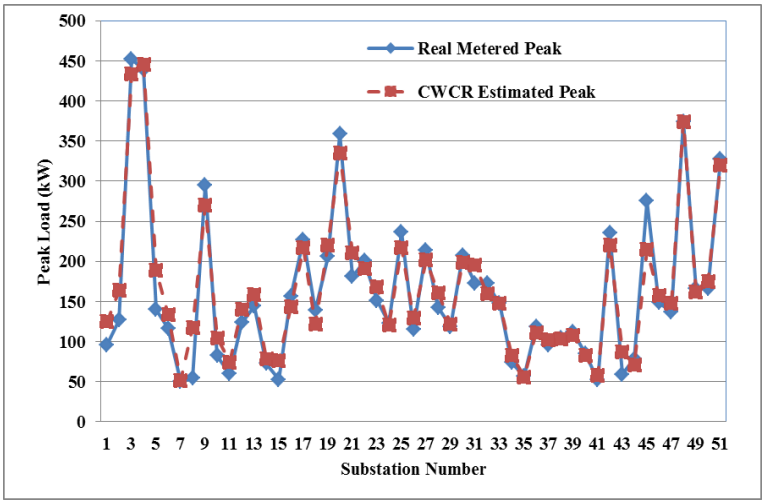

Figure 5-5 Metered and estimated peaks by CWCR

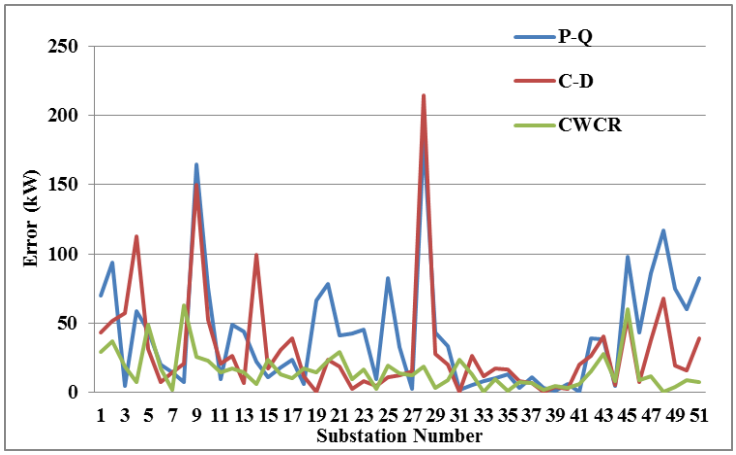

Figure 5-6 Error comparison of different methods

Table 5-1 Error Analysis of different estimation methods (kW)

	Average Absolute (kW)	Maximum (kW)	Minimum (kW)	Standard Deviation (kW)
P-Q	34.61	435.25	0.23	70.03
C-D	29.86	387.02	0.06	76.77
CWCR	14.73	179.09	0	28.63

The absolute errors between real peaks and estimations by the three methods for sample substations are plotted in Figure 5-6. The P-Q and C-D estimations produce much larger errors consistently compared with CWCR. The overall errors for all 800 substations are summarised in Table 5-1. The P-Q method produces the largest average and max errors, 34.61 kW and 435.25 kW respectively. C-D method shows slight improvement in terms of errors over P-Q method. The errors are reduced to 29.86kW and 387.02 kW respectively. The CWCR approach substantially reduces the average error to 14.73kW and the max error to 179.09 kW. Moreover, a significant drop in the error standard deviation can be seen from CWCR estimations, from 70 kW down to 28 kW. It reflects that the proposed CWCR approach performs very well over the existing approaches, producing more accurate and stable peak estimation for LV substations.

The results are also compared in terms of goodness to fit. R squared error defined in (5-8) is used to assess how a model fits the observations. The R squared error usually ranges from 0 to 1, with 1 indicating a perfect fit to the data.

$$R^2 = 1 - \frac{\sum_{j=1}^{J} (P_{jm} - P_{je})^2}{\sum_{j=1}^{J} (P_{jm} - \overline{P}_m)^2} \tag{5-8}$$

where, P_{jm} is the metered peak of the j^{th} substation; P_{je} is the estimated peak of the j^{th} substation; $\overline{P}_m$ is the mean value of all metered peaks.

For CWCR method, the R squared error is calculated for each cluster as well as for all of them to assess its performance in terms of prediction accuracy. As P-Q and C-D methods are not cluster-specified, they are only assessed in terms of overall R squared error. Table 5-2 shows a comprehensive comparison between the three methods in terms of R squared errors. The improvement of CWCR is substantial in terms of both

overall and clustered R squared errors. Clusters 8, 9 and 10 show the best performance mainly because of their typical load types. However, they may need further validation due to that their sample sizes are small (less than 30 substations). The domestic dominated clusters (2, 3, 4, 7) generally show better performance than commercial dominated clusters (1, 6) as commercial customer loads are more diverse.

Table 5-2 Goodness-of-fit Comparisons

Cluster	R Squared Error of CWCR	R Squared Error of C-D method	R Squared Error of P-Q method
1	0.84	0.59	0.49
2	0.87		
3	0.94		
4	0.79		
5	0.94		
6	0.71		
7	0.93		
8	1.00		
9	0.99		
10	1.00		
Overall	0.88	0.59	0.49

The overall R squared errors of CWCR method is 0.88 and those from C-D and P-Q methods are 0.59 and 0.49 respectively, clearly showing that CWCR has better fit over other two in estimating peak demand.

As the same substation may be clustered into different clusters in different seasons and days, the scaling coefficients are thus season-day specified. The results above are based high-summer scenario. Table 5-3 lists the average prediction accuracy in terms

of R-squared error in different seasons. It can be seen that the scaling gives very stable performance in a year.

Table 5-3 R Squared Error for all clusters and seasons

cluster	Spring	Summer	High Summer	Autumn	Winter
1	0.99	0.93	0.84	0.87	0.83
2	1.00	0.89	0.87	0.94	0.97
3	0.96	0.97	0.94	1.00	0.84
4	0.45	0.75	0.79	0.90	0.99
5	0.99	0.98	0.94	1.00	0.99
6	0.79	0.78	0.71	0.88	1.00
7	0.80	0.98	0.93	0.76	0.88
8	0.59	1.00	1.00	0.98	0.70
9	0.96	0.99	0.99	1.00	0.94
10	1.00	1.00	1.00	1.00	1.00
Average	0.93	0.90	0.88	0.91	0.85

5.6.2 Cross Validation

N-fold cross validation is adopted to prevent over-fitting in peak estimation. The whole substation set is split into N folds of the same size and each fold is used as testing set in turn. Every time, CWCR, P-Q and C-D methods are used on the training set to develop peak estimation, validated by the testing set.

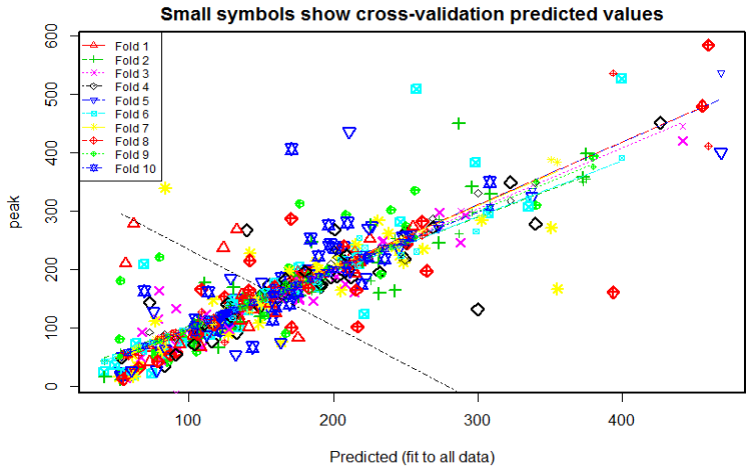

Figure 5-7 Ratio of metered and estimated peaks by P-Q method

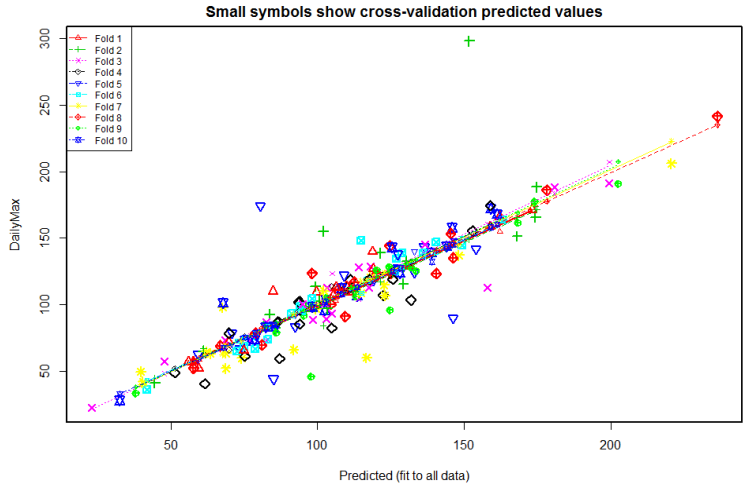

Figure 5-8 Ratio of metered and estimated peaks by cluster regression (cluster 4)

Figures 5-7 and 5-8 show the results from an example of 10-fold cross validation. The substations are divided into 10 folds, and each one is used as testing set in turn. Each color in the figures represents one turn of the process. The X-axis is the predicted value and Y-axis represents the real metered peaks. Once the ratio between the two values is closer to 1, it means the peak estimation is more accurate. It is seen that the results in Figure 5-7 (P-Q method) are a bit far from the benchmark line (slope 1). By comparison, Figure 5-8 indicates that the coefficients derived by CWCR perform well and are more stable for all 10 testing sets than those from P-Q and C-D methods.

Table 5-4 Comparison of Cross Validation on P-Q and CWCR

		MS of all data validation (kW)	Average MS in cross validation (kW)
P-Q method		2933	16281
C-D method		2532	4751
CWCR	cluster		
	1	2004	3671
	2	1836	2640
	3	490	825
	4	391	415
	5	92	316
	6	2529	4474
	7	36	184
	8	Insufficient samples	Insufficient samples
	9	Insufficient samples	Insufficient samples
	10	Insufficient samples	Insufficient samples

Table 5-4 shows the mean squares of residuals (MS) of the methods across all data validation and cross validation. All data validation indicates the model is developed and tested by using the same set of data- all 800 LV substations. MS is used to represent the mean difference between estimated peaks and real metered peaks. If a method performs well on all the testing sets through cross validation, the MS of the cross validation should be close to the MS of all data validation itself.

The P-Q method performs with a MS of 2933 in all data validation, while the average MS in 10-fold cross validation is 16281, which is considerably higher. In another word, the P-Q method produces significantly larger residuals when tested on some of the 10-fold data sets compared to testing on all sets. It indicates that the P-Q model gives unstable estimations when the training data set is different, and thus it is unlikely to be applicable to other LV substations (e.g. testing sets) or new LV substations. On the other hand, the cross validation MS of CWCR method is close to all data validation and it is very stable through all clusters. The reason is that similar types of substations are clustered together by the clustering and classification in Chapter 4, which requires a smaller sample size and provides a lower level of uncertainties. This indicates that CWCR model is less dependent on the training data and the estimation is thus more stable to be applicable to other LV substations. The 10-fold cross validation is not suitable for clusters 8, 9 and 10 due to too small sample sizes.

5.7 Discussion on the Use of Network Templates

In this section, two distribution networks are used to assess the effectiveness of the designed prefixed and dynamic dispatch strategies, demonstrating the impacts on network conditions and quantifying the benefits of applying the jointly owned energy storage. The study is first conducted on a typical settlement day, which is then extended to a year.

The developed normalised templates with the LV substation peak estimation are compiled into the classification rules introduced in Chapter 4. The combination of the work in the Chapter 4 and 5 is able to visualise the conditions of unmonitored LV substations by only using available fixed information. In order to make the templates easily used by other DNOs, an interface tool was developed to map unmonitored LV

substations only based on fixed data [74]. They are classified by their fixed information, such as customer mix and network structures, to the most similar clusters according to the probability of belonging. Benchmark load and voltage profiles from the developed templates of the highest probability are thus applicable to these substations to understand their conditions.

The profiles of extensive LV networks are reduced to 10 manageable and representative templates. They can be widely used as an effective tool and platform for different analysis conducted by DNOs, such as network planning, operation, state estimation, and demand side management, etc.

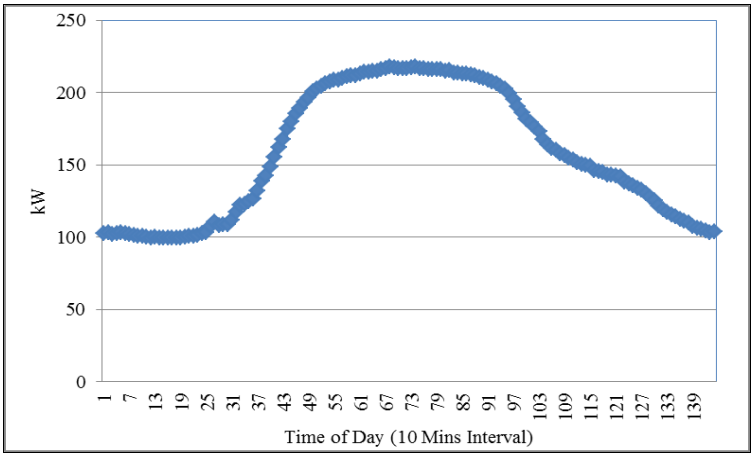

Figure 5-9 Load profile template of cluster 1

Table 5-5 Low carbon capacity for LV substations in cluster 1

Type	Comment
Workplace / retail EV charging	Unsuitable time of day pattern as need is coincident with prevailing peak
Overnight EV charging	Very suitable
Heat Pump	Only if linked with insulation or heat storage to permit off peak operation
PV	Suitable - complementary to both power and voltage curves
CHP, Hydro, Wind	Since generation is not naturally limited to time of day, potential need for constraint for voltage reasons off peak

One example of using the developed templates is to understand the capabilities of unmonitored LV substations to accommodate LCTs. The templates provide a straightforward way to quantify the available thermal "headroom" of these substations over time. Taking the load profile of template 1 in Figure 5-9 as an example, its ability to absorb the different LCTs is detailed in Table 5-5. Template 1 mainly represents load profiles of substations dominated I&C customers. As analysed in Table 5-5, these substations are not suitable for daytime EV charging. EV charging normally appears during daytime and might be coincident with the existing substation peak, producing even higher peaks. By contrast, overnight EV charging is very suitable for substations in template 1 as the exiting demand valley occurs during this period, without creating new peaks.

Table 5-6 Load factors for all clusters and seasons

cluster	Spring	Summer	High Summer	Autumn	Winter
1	0.79	0.73	0.76	0.66	0.68
2	0.96	0.86	0.85	0.74	0.82
3	0.84	0.77	0.80	0.87	0.60
4	0.72	0.72	0.71	0.61	0.76
5	0.75	0.66	0.70	0.72	0.63
6	0.65	0.64	0.65	0.69	0.91
7	0.70	0.82	0.82	0.61	0.72
8	0.66	0.98	0.98	0.74	0.63
9	0.72	0.72	0.78	0.97	0.45
10	0.42	0.35	0.41	0.53	0.63

The derived templates can also be used to assess transformer loss and aging. Table 5-6 lists the load factors for different clusters in different seasons, which are directly calculated from the normalised templates. These values can be used to calculate transformer loss based on its TLPs.

5.8 Chapter Summary

This chapter proposed an effective CWCR method to estimate the peak demand for LV substations only based on available fixed data. It develops a contribution factor to facilitate cluster-specified peak estimation. The extensive comparison demonstrates that the accuracy and stability of peak estimation has been substantially improved in terms of both R squared error and performance of cross validation.

Chapter 6

Spectral Load Profiling for Individual Customer: Feature Extraction

THIS chapter moves to load profiling for individual customers. To overcome several challenges brought by smart metering data, it is investigate to extract the data features on spectral domain

6.1 Introduction

Chapter 4 and 5 have developed the LV network templates to visualise the LV networks without extensive monitoring. The proposed three-stage method successfully cluster, classify and scale load profiles on time-series, thus extracting the smart data (meaningful information) from representative monitored area.

The recent smart meter roll-outs are providing new opportunities to understand energy patterns at a more granular level: individual household. However, the smart metering data challenges the time-series load profiling method due to their volumes, volatility, and high uncertainties. A promising alternative is to decompose time-series load data into spectral domain, where i) the irregular load profiles can be characterised by the underlying periodic spectral components, and ii) big load data can be represented by a small number of features extracted from spectral components. This chapter assesses the performance of feature extraction of two decomposition techniques: DFT and DWT. The performance is evaluated by i) load characterisation: to decompose volatile load profiles consistently; ii) data compression: the trade-offs between the accuracy of the reconstructed profile and the degree of reduction in data sizes. Assessments are performed on different granularity levels: from individual customer, averaged customer to LV substations. Results show that DWT significantly outperforms DFT for individual smart-metered customers in terms of capturing volatility with the least number of coefficients. Based on the "feature representation" in Chapter 6, Chapter 7 will propose a novel MRC technique to create customer load profiling from smart metering data, enabling effective DSR from individual and/or aggregated customers.

6.2 Problem and Proposed Solution Statement

The increasing number of LCTs, e.g. EVs, PV and heat pumps, will bring pressures to distribution networks in terms of thermal and voltage constraints [12, 33]. One solution to mitigate these pressures is DSR, where customers vary demand to price signals to support DNOs in addressing network problems [88]. The realization of DSR is commonly based on the flexibility of customers' electricity consumption, which requires characterisation of customers' load. In order to characterise the diverse and massive end customers, load profiling is traditionally used to classify customers and create TLPs.

Current TLPs in both industry and academia are inadequate to reflect individual energy usages. For the power industry in the UK, 8 TLPs developed in 1990s have been used to represent all end-customers [75], even though they have been proved to be inadequate and inaccurate [9]. In academia, due to the limited access to customer information, traditional load profiling researches are usually based on non-domestic customers, small sample sizes and averaged load profiles [22, 63, 89]. The created TLPs are used for general load research, tariff design and electricity settlements.

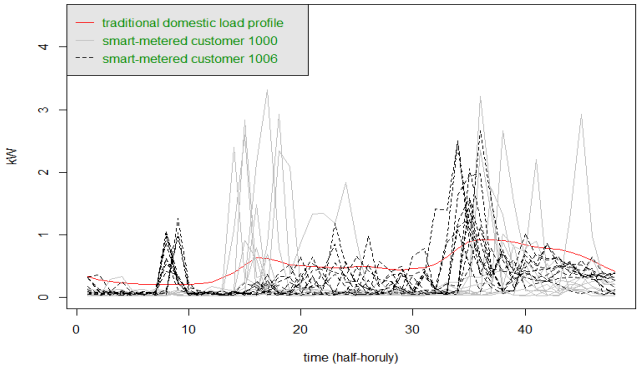

Figure 6-1 Comparison between traditional TLPs and smart metered load profiles (Data from Irish Smart Metering Project)

The characteristics of smart metered load data can be summarised as massive, volatile and uncertain. They give an insight into a more granular level of customers energy usage patterns (individual customer on individual day) while traditional TLPs are mostly developed to represent the aggregated level (average of customers or over time). The substantial differences between TLPs and smart metering data can be seen in Figure 6-1 [57]; the grey and black lines depict the smart metered load profiles of 2 domestic customers over 10 different days. In traditional load profiling, all these load profiles would be represented by a single TLP as shown by the red line in the figure, which can either represent the volatility of load profiles and/or their uncertainties between days.

Given the difference between individual smart metering data and TLPs, DSR strategies based on the TLP may not guide individual customers to effectively support energy and network needs [8, 74]. Worse, it could further aggravate energy or network problems. Smart meters now provide a new opportunity to address these issues, allowing more refined load profiles to be developed such that DSR strategies could reflect key characteristics of millions of individual customers.

However, directly applying conventional load profiling methods to smart metered data faces several challenges: i) big data require large storage space and place a heavy computation burden; ii) volatile load profiles (e.g. long-term low-demands and sudden spikes) hinder several load profiling processes including noise filtering, clustering and classification; iii) the stochastic characteristics of individual customer leads to large uncertainties in the results, e.g. the same customer belongs to different TLPs on different days.

Chapter 6 and 7 for the first time propose a novel load profiling method through capturing key features in the spectral domain. It successfully addresses the conflicts between profiling accuracy and large, volatile, and uncertain smart-meter load. Big data are compressed on different dimensions. Chapter 6 aims to reduce the number of coefficients (variables) describing each sample while Chapter 7 works on the sample size reduction. Volatility and uncertainties are addressed by a novel MRC technique proposed in next chapter. In other words, Chapter 6 focuses on "feature representation" of time-series load data on spectral domain. Chapter 7 proposes a novel "code generation" method where dominant features will be converted as "code-words" and further clustered into a "code-dictionary" for looking up.

These two chapters propose novel load profiling methods under smart metering scenarios, but more fundamentally, they provide key contributions to big data analysis. As shown in Figure 1-1, there is always a trade-off between big-data modelling errors and the number of features and clusters. For feature extraction, excess features will add noises and redundancy into data while insufficient features will lose key information. For classification, s single cluster will mix up everything while too many clusters will lead to misclassification. Previous studies usually treat them as two separate problems. This research aims to find the joint optimal number of features and clusters as the blue area shown in Figure 1-2. Chapter 6 propose a new method to extract key features, which will be innovatively used for further classification in Chapter 7.

This chapter assesses different spectral analysis techniques for smart metering data in terms of load characterisation and data compression. It decomposes the uncertain load profiles on the time-domain into consistent and meaningful features in the

spectral domain. Original load profiles can be accurately re-constructed by 60% of the spectral components, which are considered as features carrying key information.

The rest of the chapter is organised as follows: Section 6.2 reviews the spectral analysis techniques. Section 6.3 briefly introduces the data used in the assessment. Section 6.4 presents the decomposition and reconstruction approaches for both techniques. Section 6.5 proposes assessment methods for load profile characterisation and data compression. Assessment results on smart metering data are demonstrated in Section 6.6 and results different aggregated levels are compared and discussed in Section 6.7. Chapter summary is in Section 6.8.

6.3 Spectral Analysis and Data Description

Spectral analysis has been applied in different fields in power system. In load forecast, spectral analysis decomposes load data into components on different frequency levels, which can be forecasted separately [90-93]. In load characterisation, DC component can largely resemble the load factor and selected AC components are used to describe the load shape [22, 94, 95]. In data communication, end-users' load data is decomposed into different resolutions and encrypted separately in order to protect customers' privacy, which can be found in the high resolution components [96, 97]. In this chapter, different spectral analysis techniques are assessed to compress and characterise smart metering data.

1) DFT: The major steps for characterising load profiles by DFT are presented in [22]: i) evaluate the prerequisite conditions including sample rate and band-limitedness; ii) to adopt DFT to transform load profiles into the frequency domain; iii) to use inverse discrete Fourier transform (IDFT) to reconstruct load profiles as a sum of limited number of frequency components (harmonics). The results shows that average daily load profiles of customers can be precisely represented by a small set of frequency components.

2) DWT: wavelet transform has mainly been studied for short-term load forecasting (STLF) at system level. [90] emphasises the advantage of wavelet transform over DFT in that wavelet is able to capture short-duration pulse (e.g. particular event) and non-stationary features (e.g. seasonality). [93] adopts wavelet in pre-process stage to

filter noise and redundant data. [91] decomposes both load data and weather variables into low-frequencies and high-frequencies components, where low-frequencies can be precisely predicted. [92] attempts to predict high-frequencies by similar-day based neural network.

It can be seen that each technique has only been applied to a certain aggregated level as well as limited applications. With the upcoming big data from smart meters and smart grid, this chapter aims to assess the overall performance of the two techniques to feature and compress smart metering data.

The evaluation is implemented at different aggregated levels including individual customer, averaged customer and LV substations. Two sets of data respectively from smart grid and smart meter projects are assessed in this chapter. The smart grid demonstration project, Low Voltage Network Templates Project [92] is jointly commissioned by WPD in the UK. The variable data collection is on a 10-minute interval over the course of one year (2012-2013), including three-phase voltage, current and real power delivered at HV/LV substations.

The smart metering data are from the Irish smart meter trial project [57]. There are 6369 customers with half-hourly demand recorded over one and a half year (2009-2011). For LV substations, daily load profiles will be assessed. For individual customers, both monthly average and daily load profiles will be assessed.

6.4 Decomposition and Reconstruction

The decomposition process can be treated as a transformation from one function into a different set of basic functions. The basic functions of Fourier transform are sinusoids of various frequencies while wavelet transform adopts orthonormal wavelets [63, 98]. Reconstruction is basically the inverse transform; however, data can be compressed and characterised during this process.

Consider the daily load profile as a time series $\boldsymbol{s} = [s_0, s_1, \ldots s_{N-1}]$, where N is the daily sample size (N=144 for 10 minutes interval and N=48 for half-hourly). In order to compare the load decomposition on different aggregation levels, all daily load

profiles are normalised to $\boldsymbol{b} = [b_0, b_1, \ldots b_{N-1}]$ according to its maximum daily load as shown in (6-1).

$$b_n = \frac{s_n}{\max\{\boldsymbol{s}\}}, \quad n = 0, \ldots N-1 \tag{6-1}$$

6.4.1 Discrete Fourier Transform

Using DFT, $\boldsymbol{b}$ can be transformed from time domain to frequency domain. The spectrum of $\boldsymbol{b}$ is shown by (6-2)

$$B_k = \sum_{n=0}^{N-1} b_n e^{-j\frac{2\pi kn}{N}}, \; k = 0, \ldots N-1 \tag{6-2}$$

Where B_k is the frequency spectrum with magnitude of β_k and phase angle θ_k.

Using IDFT, the time series load profile $\boldsymbol{b}$ can be reconstructed by summing up the frequency components to $\boldsymbol{b}^r$:

$$b_n^r = \frac{1}{N} \sum_{k=0}^{N-1} B_k e^{j\frac{2\pi kn}{N}}, \; n = 0, \ldots N-1 \tag{6-3}$$

The complex coefficients can be merged in pair forming cosine functions with different frequencies and initial phase angles. However, a flaw was found at this point in previous research [22]. When N is an even number, the component of Nyquist frequency (k=N/2) was considered as part of DC. In fact, the Nyquist component, which is expressed in (6-4), is actually a triangular wave rather than a DC component.

$$b_{N/2}^r = \frac{\beta_{\frac{N}{2}}}{N} \cos(\pi n + \theta_{\frac{N}{2}}), \; n = 0, \ldots N-1 \tag{6-4}$$

The mistake was found to cause significant errors in our reconstruction and assessment. It is thus corrected and the reconstruction of time series *b* can be expressed by (6-5):

$$b_n^r = \frac{\beta_0}{N} + \sum_{k=1}^{\frac{N}{2}-1} \frac{2 \times \beta_k}{N} \cos\left(\frac{k 2\pi n \Delta t}{N} + \theta_k\right) + \frac{\beta_{\frac{N}{2}}}{N} \cos\left(\pi n + \theta_{\frac{N}{2}}\right)$$
$$, \; n = 0, \ldots N-1 \tag{6-5}$$

Figure 6-2 illustrate the decomposition by DFT. The volatile black line is the real load profile of a sample customer. The red line is its DC component representing the first part in (6-4). The rest colorful lines are the AC components with different frequencies. Summing up these components can get artificial time series that resemble the original load profile.

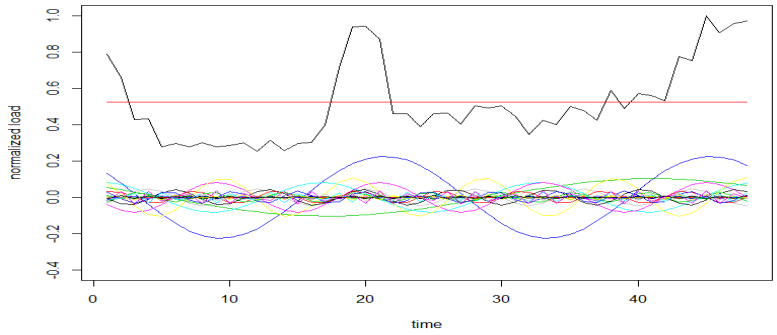

Figure 6-2 Load profile decomposition by DFT

6.4.2 Discrete Wavelet Transform

As time information is lost in Fourier transform, it is inefficient to decompose non-stationary signals, whose frequency components are varying over time. Fourier transform requires a large number of harmonics to express volatile load profiles characterised as spikes or needle peaks.

Wavelet analysis corrects the deficiency by introducing a wavelet that decays in a limited time window. It enables each component to have different scales and shifts over time. The decomposition process can be illustrated by Figure 6-3. The load profile is decomposed by high-pass and low-pass filters. The coefficients of the filters are determined by the choice of mother wavelets. The down-sampling process breaks down original load profiles into lower resolution components. Higher level of decomposition process will generate lower resolution components. The large-scale

components are called “approximation” (A) while Small-scale components are called “details” (D).

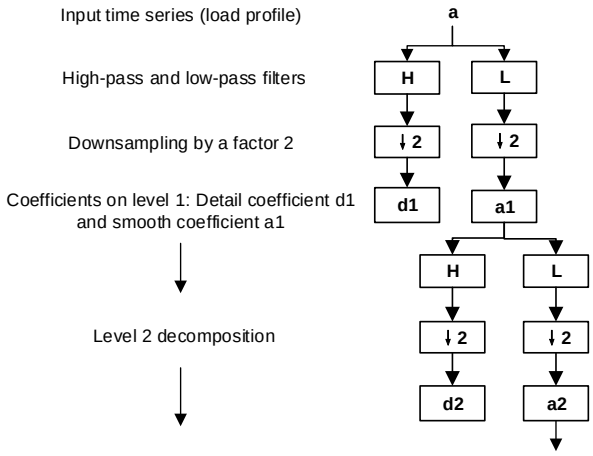

Figure 6-3 Multi-resolution analysis by DWT

The reconstruction starts from the coefficients d1 and a1. Through up-sampling and reconstruction filters, approximation component *A1* and detail component *D1* can be obtained. By this way, the original load profile can be represented by multi-resolution analysis (MRA) shown by (6-6):

$$X = D1 + A1 = D1 + D2 + A2 = \ldots = \sum_{j=1}^{J} D_j + A_J \tag{6-6}$$

where X is the reconstructed load profile; A_j and D_j are the approximation and detail components at level j, J is the total levels of decomposition.

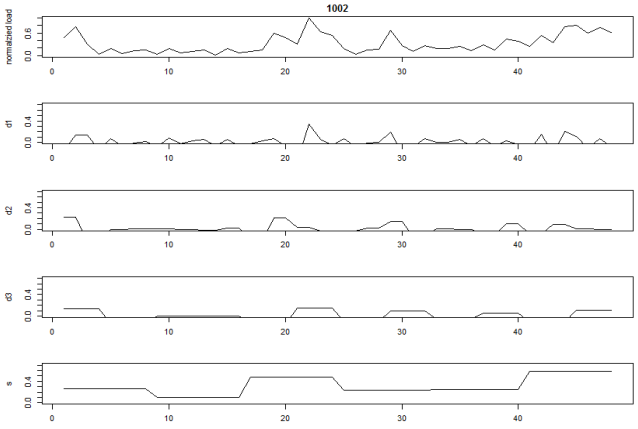

Figure 6-4 Load profile decomposition by DWT

This chapter chooses *haar* as the mother wavelet and a decomposition level of 3. *Haar* is likely to be coherent with the nature of individual customer's load profile as the square wave can better portray the turn on-off of domestic appliances. Figure 6-4 gives an example of using DWT to decompose individual customer's load profile. Curves from top to bottom are original load profile, D1, D2, D3 and A3 respectively.

6.5 Assessment Method

DFT and DWT are compared as spectral representations of load profiles on time-series. The assessment is focused on feature representation in terms of: i) load characterisation and ii) data compression.

The assessment of DFT takes the following steps: i) Data Pre-process: un-structured data sets are firstly cleaned, sense-checked (Appendix A), and organised into the same structure. Daily load profiles are normalised to certain range; ii) using DFT to decompose daily load profiles into frequency coefficients: magnitudes and phase angles of all components; iii) load characterisation: evaluate the coefficients in terms of composition, correlation and consistency (variations of the daily coefficients of the same customer over time); iv) data compression: using a limited number of components, from one to all, to represent the original load profile.

Table 6-1 DFT coefficients of a sampled load profile

Frequency	Amplitude	Phase
0 (DC)	0.72	NA
1/48	0.173	1.82
2/48	0.151	1.55
3/48	0.027	-1.73
,…	……	……
23/48	0.003	0.24

The data compression investigates the trade-off between profiling accuracy and data size reduction. It is noted that the low-frequency components, which depict the average loading level, usually dominate the magnitudes. Table 6-1 lists the DFT component of a sampled customer's load profile. As the frequency increases, the

magnitude of component dramatically drops. Aggregation of the first few DFT components is expected to capture the original load profile with high accuracy while the data size can be significantly reduced.

The representativeness of reconstructed load profiles are evaluated by the following indices: Peak Magnitude Error Index (PMEI), Maximum Magnitude Error (MME), Mean Absolute Percentage Error (MAPE), Peak Time Error (PTE). All metrics are defined in (6-7)-(6-10), where b and b^r are the original and reconstructed load profiles; $t_{max(a)}$ is the time when peak load occurs in the profile b. This chapter follows the same criteria for reconstruction assessment as in [22]. A reconstructed load profile is considered satisfactory if the PMEI, MME and MAPE are all below 5% and PTE is shorter than 2 hours.

$$PMIE = \left| \frac{\max(b) - \max(b^r)}{\max(b)} \right| \times 100\% \tag{6-7}$$

$$MME = \max\left(\left| \frac{b - b^r}{b} \right| \times 100\% \right) \tag{6-8}$$

$$MAPE = mean\left(\left| \frac{b - b^r}{b} \right| \times 100\% \right) \tag{6-9}$$

$$PTE = t_{\max(b)} - t_{\max(b^r)} \tag{6-10}$$

The assessment of DWT follows similar steps to those of DFT. However, the data reduction method for DWT is modified. Besides the use of a limited number of components, it is noted that DWT components, especially the small-scale ones, have very low magnitudes through most of the time windows. Thus, the additional method for DWT data reduction is to remove the low-demand periods of each component. Coefficients below a pre-defined threshold will be set as zeros. By this way, the number of non-zero coefficients can be significantly reduced.

6.6 Results for Smart Metering Data

This section introduces benefit quantification methods to measure the benefits that can be realised through applying the proposed operation schemes to energy storage.

6.6.1 Individual Customer

The most unique characteristic of daily load profiles of individual customer is volatility. Figure 6-5 shows the daily load profiles of customer 1002 in July 2012. The significant volatility of daily load profiles (grey) makes it inaccurate to represent them by average (red). It is also difficult to use any random day to represent the month unless some meaningful information can be extracted from these irregular load profiles.

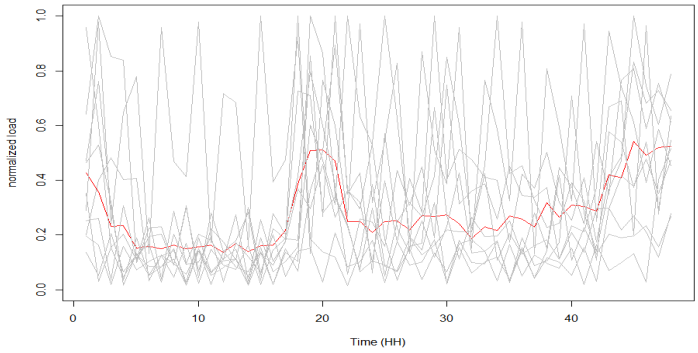

Figure 6-5 Daily load profiles of customer 1002 in July 2012

6.6.2 Load Characterisation

Theoretically, it is suggested the volatile load profiles can be decomposed into more stable and meaningful components by DWT compared with DFT. The reason is that DFT is periodic and stationary. It requires many high-frequency components to resemble the volatility of original load profiles. The shift of "needle peaks" from original load profiles may result in large variation in the DFT coefficients (phase). On the other hand, DWT is dynamic on both frequency and time domain, which enables it to capture the sudden spikes and hold the underlying trend at the same time.

In this assessment, DFT and DWT are both used to decompose the daily load profiles of 6369 customers through a year. It is found that DWT is better at load characterisation by two advantages: i) DWT decomposes the load to more meaningful components; ii) the DWT coefficients are more consistent within the same customer through days.

Figure 6-6 demonstrates the decomposition components from DWT. Each daily load profile in Figure 6-5 is decomposed by DWT into 4 components: A, d3, d2, d1, with scale from large to small. The observations are:

i) The A components describe the underlying trend of daily load profiles. For the same customer, A components are generally consistent through different days. With further classification of seasons, months and day types, the similarity will increase. As in Figure 6-6, Customer 1002 shows a fundamental usage pattern of "double-peak" in July;

ii) d3 and d2 components represent more random activities and short-interval usage (e.g. kettles). Figure 6-6 clearly sees the low-demand time from 1 a.m. to 8 a.m. (sleeping time) and busy time in the morning and evening;

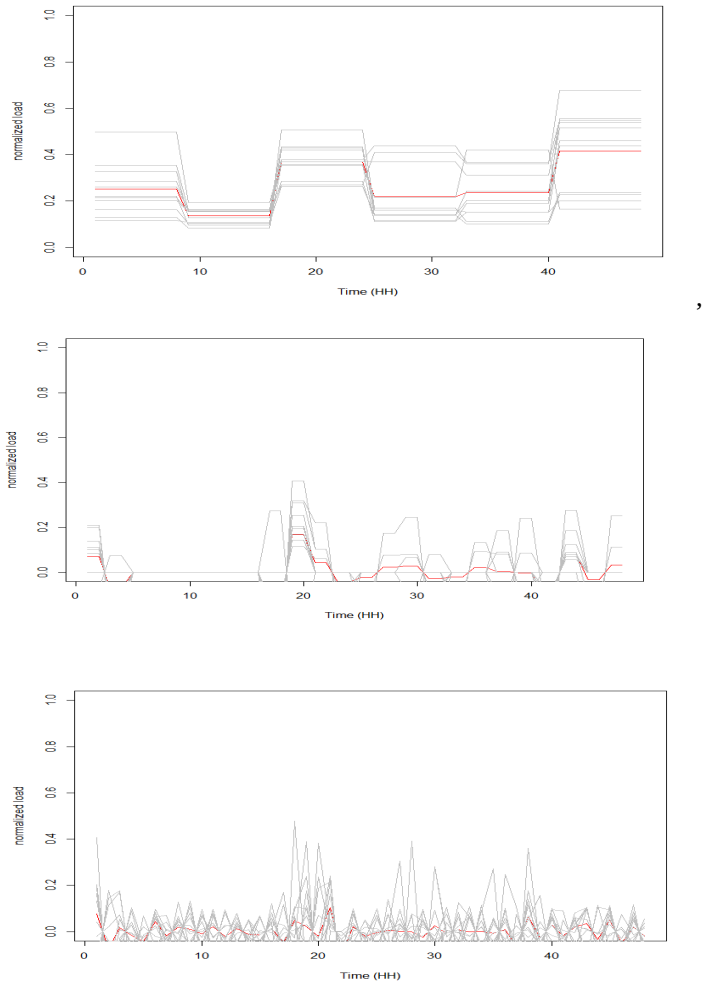

Figure 6-6 Decomposition components from DWT customer 1002

iii) d1 component has the smallest scale. It contains random spikes which are possibly caused by the turn-on of some appliances. It is also noted that some of the d1 components are quite periodical, likely to represent white goods such as refrigerators.

In contrast, the periodical sinusoidal components from DFT reveal less information as shown in Figure 6-7.

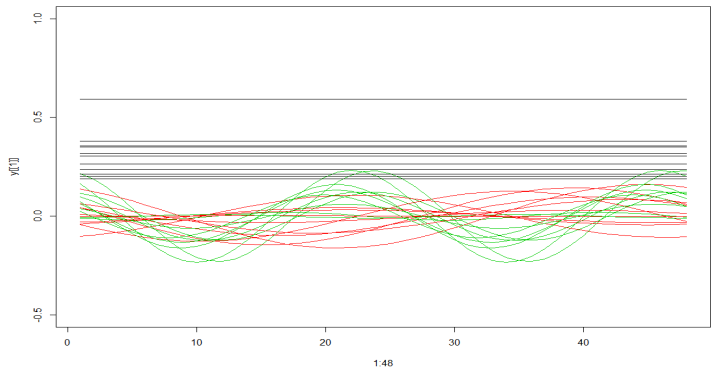

Figure 6-7 Periodical sinusoidal components from DFT

6.6.3 Data Compression

A reduced number of the transformed coefficients can be used to re-construct the original load profile with errors. Another assessment is to evaluate the trade-off between representativeness of the reconstructed load profile and data reduction. The idea is based on the assumption that the reconstruction is dominated by low-frequency components. Using the first few coefficients will adequately resemble the original load profiles as they preserve the majority of the spectral energy, which is calculated as the sum squares of coefficients' magnitudes.

The assumption is verified by test on all load profiles, reconstructing from low to high frequencies by both DFT and DWT. Figure 6-8 demonstrates the accumulated energy by keeping different number of coefficients from DFT and DWT. As shown in the figure, keeping all 48 coefficients, both methods will preserve 100% of the energy in original load profiles while the first coefficient alone contains 20% original energy. The observations are:

i) The first coefficient of both methods has around 24% energy of the original load

profile, which is consistently close with the load factor (average/peak) of the original load profile. It is expected because the DC component usually stands the mean value of original signal, and in our case (normalised load profile with peak "1") the load factor is exactly the mean;

ii) the large-scale component of DWT contains more energy, with over 99% energy after first 6 coefficients. The energy spread more evenly on DFT coefficients, reaching only 90% after 24 coefficients. It shows that with the same data reduction, DWT reconstruction will preserve more energy of the original load profile.

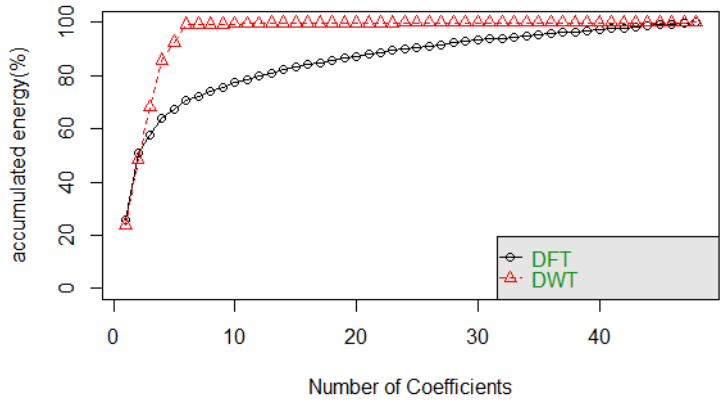

Figure 6-8 Accumulated energy by keeping different number of coefficients

The data size of DWT can be further reduced by eliminating all "near-zero" coefficients. Especially the small-scale components of DWT, which are likely to see low-demand for long time and only several spikes over a day, contains many coefficients close to zero. Eliminating those coefficients will hardly affect the reconstruction accuracy meanwhile reducing the data size considerably.

To further compare the data reduction ability of DFT and DWT, an extensive comparison is conducted between original load profiles and reconstructed load profiles. Four indices (PMEI, MME, MAPE and PTE) widely used in load profiling are adopted here.

6369 customers' daily load profiles are reconstructed with different sizes of reduced data. The test is to find the minimum data size required to meet the reconstruction accuracy. In other words, the aim is to find the possible maximum data reduction

while keeping the reconstruction error under the threshold. In this chapter, the error threshold is set to be 5% for PMEI, MME, MAPE and 2 hour for PTE. It follows the previous studies so that the results are comparable.

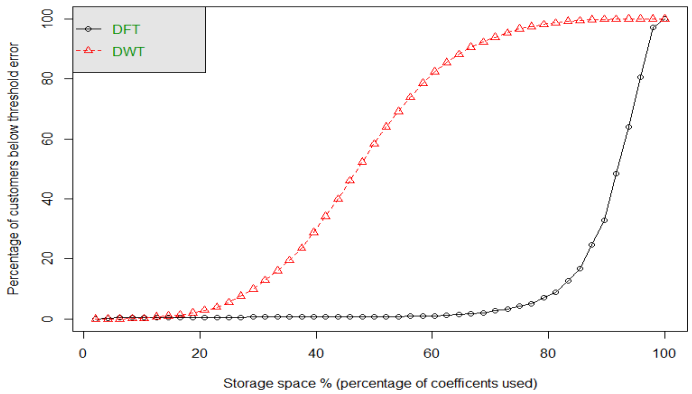

Figure 6-9 Percentage of customers who can be reconstructed under the threshold error with different data size

Figure 6-9 shows the percentage of customers who can be reconstructed under the threshold error with different data size. The x-axis is the data size (100%=48 coefficients) used to reconstruct the load profiles. For example, using half of the DFT coefficients, only 0.8% of the total customers' load profiles (about 46 customers) can be reconstructed with an error below threshold. However, using half of the DWT coefficients, 58% of the customers' load profiles can be satisfactorily reconstructed. Other main findings are:

i) The reconstruction can hardly meet the accuracy requirements with less than 20% of the coefficients for both techniques. The pass rate starts to increase when using more than 20% of DWT data. However, the DFT pass rate remains low until using more than 80% of its coefficients. For volatile load profiles, DFT needs relatively complete high-frequent component sets to resemble the sudden spikes while DWT can handle that with only a few small-scale coefficients.

ii) even with all of the DFT coefficients below Nyquist frequency (47/48), still 2.8% (174 out of 6369) of the customers' load profiles cannot be reconstructed below the threshold error. However, with 47 of the DWT coefficients, all load profiles can be successfully recovered.

iii) the largest gap between the 2 techniques occurs at 75% of the data size. Using

75% of DWT coefficients can recover 96.7% of the original load profiles. However, only 4.2% of the original load profiles are recovered by 75% DFT coefficients. The difference is as high as 92.5%. The fundamental reason is that the natural shapes of smart metering load profiles are more coherent with Haar wavelet than sinusoidal waves.

The benefits, in terms of savings in both network investment deferral and energy cost are considered.

6.7 Assessment over Different Aggregation Levels

Different applications of load profiles focus on different aggregation levels. Some tariff design is based on aggregation over time while network planning pays more attention to aggregation over customers. We roll out similar assessments as in 6.6, but on different aggregation levels. For aggregation over time, monthly average load profiles of 6369 smart metering customers are tested. For aggregation over customers, the daily load profiles from 800 LV substations are used.

6.7.1 Monthly Averaged Load Profiles

Figure 6-10 and 6-11 show the difference between monthly average and daily individual load profiles. The black line in Figure 6-10 is the daily load profile of customer ID 1000 on 1st July 2012. The red line is the reconstructed load profile by the first 3 DFT components (6 coefficients). The blue line is the reconstructed load profile by the largest 6 DWT coefficients. Clearly, with the same data size, reconstruction of DWT is much better than that of DFT.

In Figure 6-11, the black line is the average load profile of customer ID 1000 in July. It is smoother than the daily load profile. Using the same reduced data size to reconstruct the average load profile, DFT shows a much better performance compared with that on daily load. Although DWT still resemble the original load profiles better than DFT, the gap is substantially narrowed. This is also illustrated by Figure 6-12, which is a comparable plot to Figure 6-9. It is the successful reconstruction rate for monthly average load profiles with different data size. The performance of DWT is very similar with that of daily load profiles. However, DFT shows an overall

improvement. Using 80% of the DFT coefficients can recover 5.7% of the daily load profiles, but 48.5% of the monthly average load profiles.

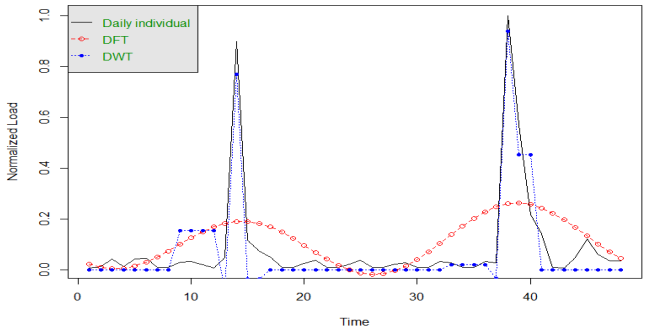

Figure 6-10 Daily individual load profile and reconstructions by reduced DFT and DWT coefficients

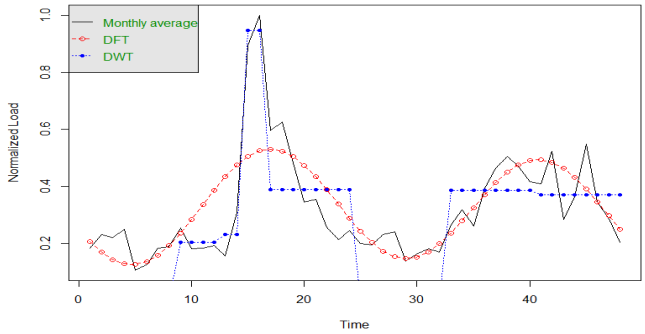

Figure 6-11 Monthly average load profile and reconstructions by reduced DFT and DWT coefficients

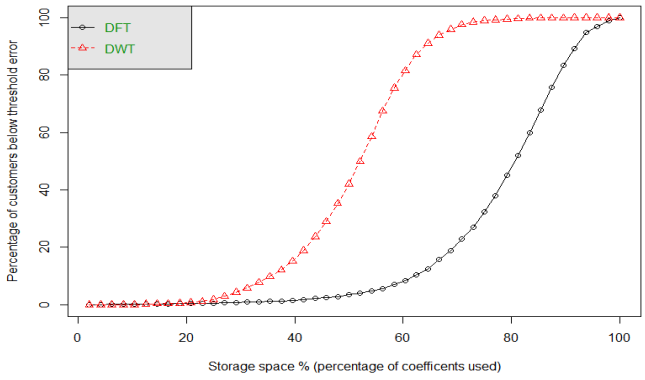

Figure 6-12 Percentage of customers who can be reconstructed under the threshold error with different data size (monthly)

6.7.2 LV Substation

The daily load profiles at LV substation are representatives for aggregated load over customers. The assessment shows that when the load profiles are granular, DWT constantly performs better at data reduction; however, DFT improves significantly as the aggregation level increases.

Substations are assessed by their customer sizes. In order to demonstrate a continuous change of customer size, some individual customers load profiles are added onto the substation artificially. Figure 6-13 shows the average minimum data required to reconstruct load profiles of different customer groups. As the customer size increases, load profiles are more aggregated and smooth. Naturally, the data required to reconstruct the load profiles decrease for both techniques. The interesting findings are:

i) When the customer size is small, DWT is generally superior to DFT. However, when the customer size increases over 400, DFT requires fewer coefficients than DWT in terms of reconstruction.

ii) Further, when the customer size is larger than 450, the DFT steadily requires only first 3 components (12% of the coefficients) to fulfil the reconstruction. This figure further becomes constant when over 700 customers.

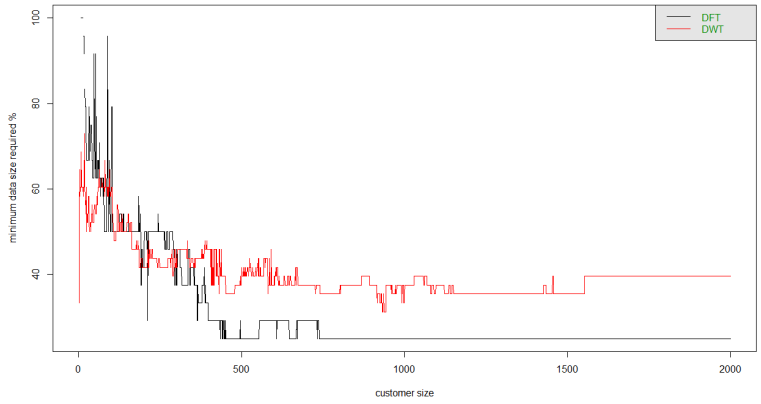

Figure 6-13 Average minimum data required to reconstruct load profiles from DFT and DWT coefficients for different customer groups (PMEI, MME, MAPE< 5% and PTE<2 hours)

iii) In contrast, DWT averagely requires 48% of the data to reconstruct small group of customers' load profiles. For larger group, it averagely requires 17 (35%) of its

coefficients with some fluctuations.

6.8 Chapter Summary

The spectral analysis provides an opportunity to analyse load profiles in spectral domain. The direct advantage includes data reduction and load characterisation, especially for smart metering data, which are extremely volatile, irregular and massive.

This chapter presents a comprehensive review and assessment on two spectral analysis techniques including two extreme cases. In DWT, Haar is chosen as the most compact wavelet while sinusoidal wave in DFT gives a global support. A new data reduction and load characterisation method based on DWT is proposed for load profiles of smart metering customers.

Based on the result, DFT could be effective for load profiling at high-aggregated level while DWT is more promising at granular level. The assessment could be useful for various smart-meter related applications such as smart tariff design, DSR and customer classification. Chapter 6 extracts key features of smart metering on spectral domain. Based on these discriminated coefficients, nest chapter will propose a novel multi-resolution load profiling method to classify loads from key features.

Chapter 7

Spectral Load Profiling for Individual Customer: Multi-resolution Clustering

THIS chapter evaluates domestic demand shifting in response to smart variable tariffs. The value of it is quantified as an equivalent storage capacity for the investigation of complementarity between technical and social interventions.

7.1 Introduction

Chapter 6 has assessed different spectral techniques to extract the features of the load data. For smart metering data, the DWT coefficients can better represent the features than DFT given the same data reduction rate. It also decomposes smart metering load data into spectral domain, which enables a multi-resolution analysis of the original load profile.

This chapter proposes a novel MRC aiming to classify customers from these extracted features. It addresses three main limitations when directly applying time-series load profiling to smart metering data: big data, volatility and uncertainty. For big data, although the coefficients of each sample are reduced in part I, on the other dimension, the large sample size remains a challenge for load profile clustering. For volatility, the interferences between different factors (e.g. magnitude, overall trend, spikes and etc.) need addressing in load profile clustering process. For uncertainties, the same customer may have very different load profiles between days, which make the classification difficult.

The proposed MRC method can separates different load characteristics to different resolution levels and operates clustering analysis on each resolution level independently. The overall TLP is derived by aggregating sub-TLP on each resolution level. Thus, instead of one fixed TLP per cluster, different combinations of sub-TLPs provide a flexible way to express variances within cluster. By $\sum_{i=1}^{I} n_i$ levels of computation, MRC can express equivalent to $\prod_{i=1}^{I} n_i$ levels load profiling. Thus it addresses the three key limitations in load profiling: i) a two-stage clustering method is implemented in MRC with Gaussian mixture model (GMM) and X-means to further reduce the input size with minimum information loss; ii) it avoids the interferences from volatility by decompose the load characteristics; iii) as the GMM can give a probabilistic cluster membership instead of a deterministic one, an additive classification model based on posterior probability is proposed to reflect the uncertainty between days. The method is implemented on over 6369 smart metered customers from Ireland, and compared with industry load profiling and traditional K-

means clustering. The results show great improvement in load profiling in terms of computation storage (speed), load profile accuracy and classification flexibility.

7.2 Problem and Proposed Solution Statement

Load profiling has been widely used to efficiently represent various end customers, which includes group customers with similar load profiles into classes and identify their TLPs. The popularization of smart meters brings the opportunity for more accurate load profiling. It can provide supports for demand side responses (DSR), customer load forecast [8, 23, 74, 95], low carbon network planning [74, 92] and smart tariff design [88, 94].

In Chapter 6, Spectral analysis techniques have been assessed to successfully resolve the issues of load profile characterisations and data reductions on frequency domain. Based on the knowledge, this chapter focuses on the load profiling of individual customers. Fundamentally, it will cluster and classify the features extracted in Chapter 6.

The state-of-art for customer load profiling generally follows the two-stage "clustering and classification" process reviewed in Chapter 2. Many different clustering techniques have been adopted in literature including: K-means [63], hierarchical [8, 19], fuzzy-c-means [71], SOM [72]. A comparative review and assessment of different techniques can be found in [89]. However, directly applying these clustering techniques on load data from smart meters has three major limitations.

i) Big data: it increases computational and storage burden for clustering analysis. Although feature selection techniques including principal component analysis (PCA), Sammon map and curvilinear component analysis (CCA) has been assessed in [89] to reduce the size of input data to clustering process, there are two main drawbacks: i) they only reduce the number of variables of each sample, but not the massive sample size on the other dimension; ii) Also they discard some sample points, which cannot be recovered, thus causing detailed information loss. For example, the common evening peak existing in most load profiles are likely to be removed by PCA, which cannot recover the original profiles.

ii) Volatility: Most of the previous researches only concentrate on non-residential customers or average load profiles because daily load profiles of residential customers are extremely volatile. A sudden spike or a tiny time shift (communication delay) may lead to completely different clustering results; for instance, two very similar (or identical) load profiles, which are both volatile but one with slight delay, will be clustered into different groups by time-series clustering. Due to the slight delay and significant volatility of two load profiles, the peak of one load profile keeps meeting the trough of the other one at each sample point, which dramatically increases the distance between two load profiles. Different factors, such as magnitudes, overall trends and spikes will interfere with each other during the clustering.

iii) Uncertainty: the same customer may have very different load profiles between days. As a result, different days may have different cluster membership, which makes it difficult for customer classification. Most researches [22, 89] used the averaged load profile over a time span (e.g. monthly) to smooth individual daily load profiles and also to prevent uncertainties in customer classification. However, averaged load profiles can be very different from individual ones (non-convex sets) and some important details from smart meters (e.g. particular event, sudden spike and empty house) are lost by averaging load profiles.

This chapter proposes a novel MRC method. By clustering load profiles in the spectral-domain instead of time-domain, it for the first time classifies electricity end-customers directly from massive, volatile and uncertain smart metered load data. The method includes three main steps:

i) Decomposition: wavelet analysis decomposes large volume of load data into compressed coefficients on spectral domain.

ii) Clustering: GMM is adopted in the clustering process to reflect the uncertainties. Clustering on each resolution level not only breaks down the computation burden, but also isolates different load characteristics such as magnitudes, overall trends, and sudden spikes. On each resolution level, a sub-TLP is developed to represent common characteristics.

iii) Reconstruction: a classification technique is developed to allocate sampled load profile to the most-likely sub-TLP on each level. Given the wavelets orthogonal, an additive model can be used to reconstruct overall TLP by aggregating the sub-TLP on each resolution level.

The rest of the chapter is constructed as follows. Section 7.3 briefly introduces the use of GMM and X-means techniques in this study. Section 7.4 proposes the MRC method and classification model. Section 7.5 implements the method on the Irish smart metered data. Results are demonstrated and compared with other clustering methods in Section 7.6. Conclusions are drawn in Section 7.7.

7.3 Clustering Techniques

This sector briefly introduces two clustering techniques that are used in the development of MRC. Only the related theoretical backgrounds are provided in this section.

7.3.1 GMM

Mixture model develops a mixture probability density function (PDF) for observations with latent classes. The mixture PDF is described as a weighted sum of finite known PDFs. In this chapter, Gaussian distribution is adopted with several practical constraints. It is chosen because it is widely used to represent load characteristics and its simplicity of modification. Mixtures with non-normal components can also be implemented by this method [99].

Suppose $\boldsymbol{s}_j$ is the j^{th} load profile of a sample customer, and $\boldsymbol{s}_j^{(l)}$ is its multi-resolution analysis (MRA) component at (l) level. The dimension of $\boldsymbol{s}_j^{(l)}$ is $N(l)$. $f(\boldsymbol{s}_j^{(l)};\Psi)$ denotes the PDF of dependent variable $\boldsymbol{s}_j^{(l)}$. A K-component finite mixture PDF can be written as (7-1), assuming component densities $f_k(\boldsymbol{s}_j^{(l)};\theta_k)$ are specified to belong to some parametric family.

$$f(\boldsymbol{s}_j^{(l)};\Psi)=\sum_{k=1}^{K}\lambda_k\, f_k(\boldsymbol{s}_j^{(l)};\theta_k) \tag{7-1}$$

Where K is the number of mixture components, λ_k is the weight of the k[th] component. As the total probability of each component density as well as the mixture density equals to 1, the summation of all weights must be unity as in (7-2)

$$\sum_{k=1}^{K} \lambda_k = 1; \quad 0 \le \lambda_k \le 1 \tag{7-2}$$

Vector Ψ denotes all the unknown parameters in the mixture model including weights and parameters of component PDF, as $\Psi = \{\lambda_k, \theta_k\}_{k=1}^{K}$.

In GMM, θ_k can be represented by: $\theta_k = (\mu_k, \Sigma_k) \quad (k = 1,2,\ldots K)$ for multivariate case. The component PDF in (7-1) can be written as a normal distribution in (7-3)

$$f_k(\boldsymbol{s}_j^{(l)}; \theta_k) = (2\pi)^{-\frac{N}{2}} \det(\Sigma_k)^{-\frac{1}{2}} \exp\left[\frac{-(\boldsymbol{s}_j^{(l)} - \mu_k)^{\mathrm{T}} \Sigma_k^{-1} (\boldsymbol{s}_j^{(l)} - \mu_k)}{2}\right] \tag{7-3}$$

7.3.2 Parameter estimation by EM algorithm

The parameters of λ_k μ_k Σ_k can be estimated by Expectation Maximum (EM). For given J observations, the log-likelihood function of parameters can be expressed by:

$$\ln L(\boldsymbol{s}_1^{(l)} \ldots \boldsymbol{s}_J^{(l)}) = \sum_{j=1}^{J} \ln(f(\boldsymbol{s}_j^{(l)}; \Psi)) \tag{7-4}$$

Parameters Ψ can be estimated by maximising equation (7-4) using EM algorithm. In the E stage, once the parameters are estimated, each observation can be assigned to each cluster *k* by Bayes rule. The posterior probability of observation *j* belonging to cluster *k* is given by (7-5). Observation will be assigned to the cluster with highest probability.

$$P(k \mid \boldsymbol{s}_j^{(l)}) = \frac{\lambda_k f_k(\boldsymbol{s}_j^{(l)}; \theta_k)}{\sum_{r=1}^{K} \lambda_r f_r(\boldsymbol{s}_j^{(l)}; \theta_r)} \tag{7-5}$$

In the M stage, the parameters are derived by maximising (7-4) again under new posterior probabilities. The updated parameters are obtained by (7-6)-(7-8). The detailed algorithm can be found in [100]:

$$\overline{\lambda_k} = \frac{1}{J}\sum_{j=1}^{J} P(k \mid \boldsymbol{s}_j^{(l)}) \tag{7-6}$$

$$\mu_k = \frac{\sum_{j=1}^{J} s_j \times p_{jk}}{\sum_{j=1}^{J} p_{jk}} \tag{7-7}$$

$$\Sigma_k = \frac{\sum_{j=1}^{J} p_{jk}(s_j - \mu_k)(s_j - \mu_k)^{\mathrm{T}}}{\sum_{j=1}^{J} p_{jk}} \tag{7-8}$$

The EM algorithm can maximise (7-4) by following iteration steps:

1. Before the first iteration, s=0, initialising the number of clusters K (J/K>I), and a starting partition p_{jk}^{0}. This can be achieved by random partition or clustering techniques.

2. Given any p_{jk}^{s}, s=0, 1, 2 …, parameters Ψ^{s} can be estimated to maximise (7-4) under p_{jk}^{s}, $\max(\ln L)^{s}$.

3. Once parameters Ψ^{s} from s iteration are estimated, each observation can be re-assigned to each cluster by new posterior probabilities p_{jk}^{s+1}.

4. Repeat step 2 and 3 until converge: $\left|\max(\ln L)^{s+1} - \max(\ln L)^{s}\right|$ is smaller than stopping criterion.

7.3.3 X-means

X-Means clustering is very similar to conventional K-means clustering. Instead of using a pre-defined number of clusters *K*, X-means will search in a range of different values of *K*, and determine the optimum *K* based on a model selection criterion such

as BIC. The detailed algorithm is introduced in [89]. The strategy is as follows: Firstly, to assess each *K*, the K-means clustering is applied to give deterministic centroids and its BIC value. Secondly, a new centroid is introduced by splitting some existing centroids into two. This is achieved by a local K-means (*K*=2) within the centroid subset. Whether the split is meaningful is determined by local BIC improvement. The process is iterated till *K* meets its upper bound. The selection of K is determined by their BIC scores. The BIC brings a penalty term for the number of parameters in the model in order to assess the trade-off between likelihood and number of clusters. It is expressed in (7-9) to find the optimum number of clusters [101].

$$L_{BIC}(\boldsymbol{s}_1^{(l)} \ldots \boldsymbol{s}_J^{(l)}) = -2\ln L(\boldsymbol{s}_1^{(l)} \ldots \boldsymbol{s}_J^{(l)}) + J\ln(K * N(l)) \tag{7-9}$$

Where $K * N(l)$ is the number of parameters, for K clusters and $N(l)$ is the dimension of $\boldsymbol{s}_j^{(l)}$. J is the sample size.

7.4 Multi-resolution Clustering

This sector will firstly review the limitations of time-series clustering, based on which a novel MRC will be proposed on spectral domain.

7.4.1 Time-series clustering

The general steps of traditional load profiling are shown in Figure 7-1. The input data of various clustering techniques are usually time-series load profiles. It aims to group *M* load profiles into K clusters. The partition is based on metrics between load profiles (e.g. Euclidean distance).

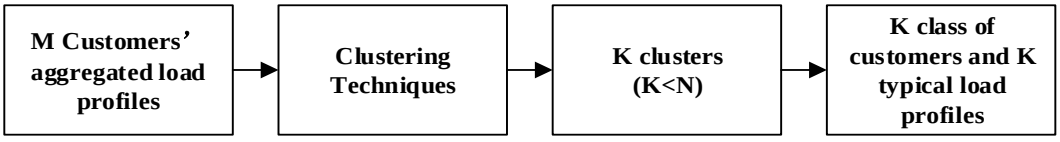

Figure 7-1 Conventional load profile clustering process

The time-series clustering shows limitations in handling magnitude, volatility and uncertainty, which have been explained in the Introduction part. Figure 7-2 depicts an example of the magnitude difference within groups. They are from real load data

clustered by K-means. As the input load data are normalised, clustering process is entirely based on the similarity of shape. Two load profiles in Figure 7-2 are clustered into the same group due to similar load shapes. However, there is a substantial difference between their original magnitudes. Therefore, the TLP of this clusters only represent load shape but not magnitudes.

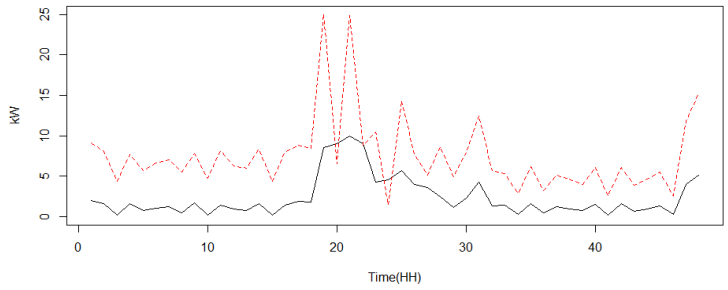

Figure 7-2 Problems with time-series clustering: magnitude difference within clusters

Due to the nature of time-series metrics, volatility has a severe impact on clustering. For instance, two very similar load profiles are clustered into different groups by time-series clustering in Figure 7-3. Due to the slight delay and significant volatility of two load profiles, the peak of one load profile keep meeting the trough of the other one at each sample point, which dramatically increases the distance between two load profiles.

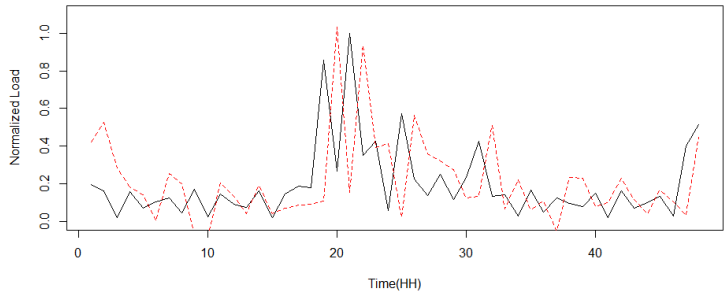

Figure 7-3 Problems with time-series clustering: time difference in spikes

As far as the literature reviewed, no research has really touched the load profiling of individual customers on individual days. One reason is the uncertainty of customer between days, which means the same customer may have very different load profiles

between days. Most studies are specified for a season or day types (e.g. weekday and weekend).

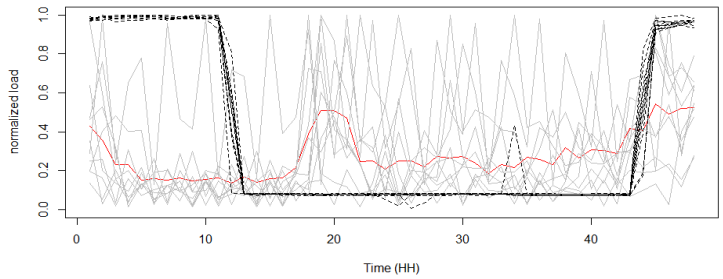

Figure 7-4 Problems with time-series clustering: uncertainties between days

However, if the daily load profiles are non-convex, the averaged load profiles will be out of the set, which will not only lose detailed information, but also cause mis-classification. In Figure 7-4, the grey lines and black lines are load profiles of the same customer on different days. A clear difference can be seen between days. The averaged load profile (red) can express days neither in grey nor black, making the TLP of this cluster much less representative.

7.4.2 MRC

A novel MRC method is proposed to cluster daily load profile on spectral domain. Based on the analysis of Chapter 6, time-series load profiles from smart meters can be successfully decomposed into spectral components, each representing different load characteristics. Multi-resolution analysis based on wavelet transfer shows an outstanding performance on individual's level in terms of capturing overall trend, isolating spikes and reducing data size. It decomposes volatile and irregular load profile into a smooth large scale component describing the underlying shape, and several small scale components describing volatilities.

The methodology of MRC is illustrated by the flow chart in Figure 7-5. It consists of three main stages: spectral analysis, clustering and reconstruction. By spectral analysis, a load profile can be decomposed into several components including magnitude, approximation (A) component, details (from D1 to Di) components. The basic idea of MRC is to implement clustering analysis on each of the component

separately, and develop a typical component (TC) for each cluster of each component level (e.g. A-1 indicates the cluster 1 on Approximation level). Accordingly, in classification process, a customer's daily load profile will have a cluster membership on every component level. The TLP is developed by aggregating assigned TCs as shown in (7-10).

$$TLP = TC_{mag} \times (TC_S + \sum_{1}^{i} TC_{Di}) \tag{7-10}$$

Where the synthesis of TCs from MRA (A to D_i levels) provides the shape of the load profile, and the magnitude TC will scale up the shape to the typical loading level.

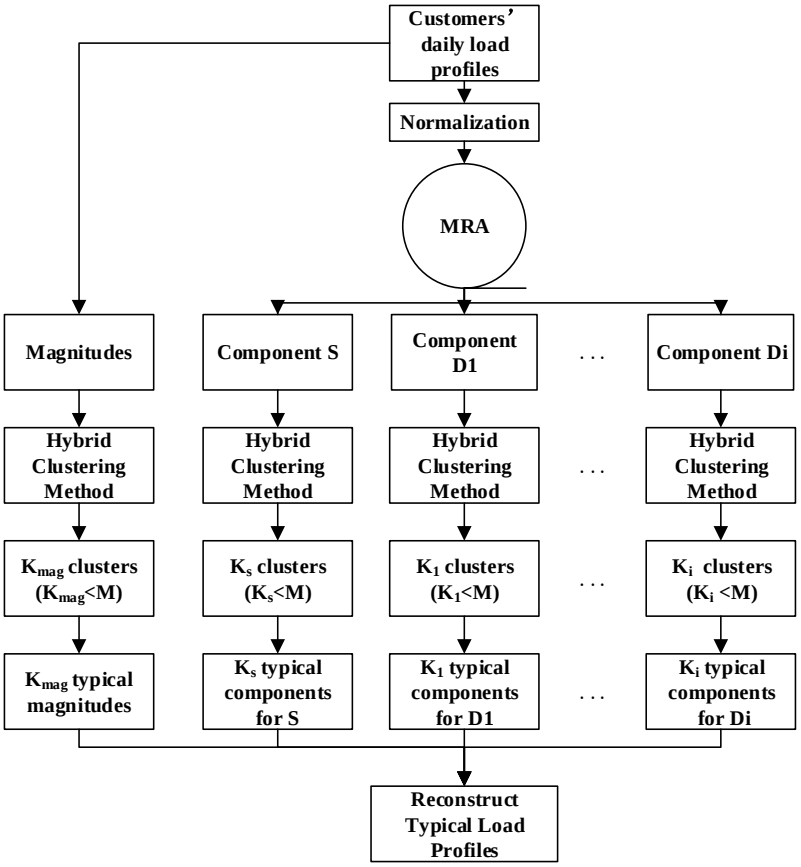

Figure 7-5 Overall methodology of multi-resolution clustering

The obvious advantage is that each clustering will focus on one characteristic without interference between each other. Other improvements include: i) as assessed in part I, input data size is substantially reduced by the spectral analysis; ii) load magnitude and shape are separately clustered, but jointly integrated in TLPs; iii) The problem caused by volatility can be resolved as overall trends and spikes are separated; iv) provide

flexibility for the number of clusters. In traditional clustering methods, due to the uncertainties, the load profiles between days can be very different, which requires a huge number of clusters to express. MRC provides an opportunity to have different cluster numbers on different levels. For example, a few clusters may be sufficient for A level as the overall trends of customers are likely to be similar and stable over days while detail levels may require more clusters to distinguish random spikes.

Although the multi-resolution analysis (MRA) in part-I reduces the data size of each sample, another issue is the number of samples. For example, in our case, 6369 customers daily load profiles over a year produces over 2.2 million observations, which requires a large memory to build up its distance matrix (for some methods, e.g. hierarchical clustering, nearest neighbour classifiers and multi-dimension scaling) . Even if the number observations can be reduced by separating season, month or day type, it is still either a heavy computation burden or a compromise on accuracy. In order to i) further express the uncertainty between days and ii) reduce the number of observations in clustering, a novel clustering method is proposed. The hybrid clustering stage shown in Figure 7-5 is expanded as the flow chart in Figure 7-6. It uses GMM to pre-cluster each customer through multiple days so that one customer's daily profiles can be represented by several typical models. Then X-means clustering will be performed only on these models. By the two-stage clustering, massive inputs are reduced into a small number of PDFs while original information is maintained in the models.

i) GMM pre-clustering: For level (*l)* of components (e.g. component s), instead of clustering all customers all days, a pre-clustering is conducted on each customer. For a customer with *J* days' load profiles $\boldsymbol{s}_j^{(l)}, j = 1 \ldots J$, GMM is used to cluster them into *K* models. Each model will be represented by a within-customer typical component (WTC), which is defined as the daily load profile with lowest uncertainty (highest posterior probability) in the cluster. The uncertainty is defined as 1 minus the max posterior probability. For the number of WTCs, i.e. number of clusters, starting with K =1, the aim is to find the least number of *K* while keeping the uncertainty of every day below threshold β. It ensures every daily component is sufficiently close to the model

centre. Thus, *J* daily components can be reduced and represented by *K* within-customer WTCs as shown in (7-11)

$$WTC_k = \arg\max_{s_j^{(l)}}(P(k \mid s_j^{(l)})), k = 1, 2, \ldots K \tag{7-11}$$

Where WTC_k is the WTC of cluster *k*, $P(k \mid \mathbf{s}_j^{(l)})$ is the posterior probability in (7-5).

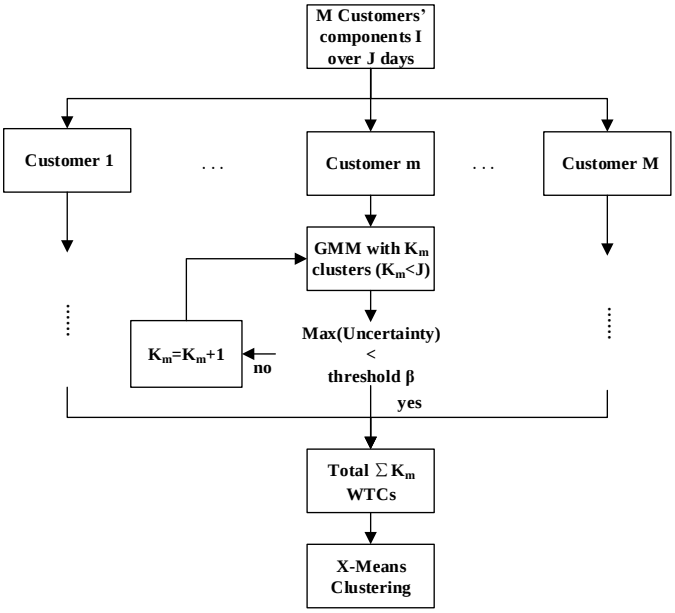

Figure 7-6 Two-stage GMM and X-means clustering implemented in MRC

Also, for each customer, the probability of every daily component belonging to each WTC is calculated for customer classification. A customer could be assigned with probabilistic classifications rather than a deterministic one. The final updated weight calculated in (7-6) is taken as the classification weight of each WTC.

ii) X-Means clustering: in the second stage, the WTCs of each customer are used as the input data of clustering. A total number of $\sum_{m=1}^{M} K_m$ WTCs are clustered by X-Means clustering. The outputs are the centers of each cluster, which are

regarded as typical components (TCs). At this stage, the input data have already been processed with data reduction, characteristics isolation and uncertainties identification; therefore, most clustering techniques could theoretically deliver decent clustering performance. However, as the number of observations can be still large especially for small-scale components, main considerations of choosing clustering techniques at this stage are: i) low computation complexity to process large sample size; ii) requiring no pre-knowledge on the number of clusters as the range could be very large. X-Means clustering, as an extended K-Means, is chosen because it inherits the simplicity of K-Means while automatically searching optimal number of clusters based on BIC.

7.5 Classifications

The classification process is to determine the cluster membership of a customer. For new customers, sometimes it requires to classify customer entirely based on only their eco-social information (fixed data) due to the limited availability of metering data. In this chapter, under the smart meter scenario, classification is based on historical sampled load data.

As the GMM pre-clustering stage group a customer's load profiles over days into several Gaussian models (the centres are the WTCs), the second stage is actually to cluster these models. After second stage of X-means clustering, all these models are clustered into Q new clusters. Each cluster q (q=1…Q) can be treated as a new mixture model, made up of different Gaussian distributions (WTCs) from the first stage. The parameters of individual PDF will not change but the weight in the new mixture model requires a normalization to ensure the unity sum. For cluster q, containing $n(q)$ WTCs (Gaussian models), the new weights are calculated as (7-12):

$$\lambda_k^q = \frac{\lambda_k}{\sum_{k=1}^{n(q)} \lambda_k}, k = 1,2 \ldots n(q) \tag{7-12}$$

Where λ_k is the weight calculated in (7-6); λ_k^q is the new weight of each component in cluster q.

As each cluster *q* is expressed as a mixture model now with new weight λ_k^q and a new class label ω_q, the classification of sampled data $\boldsymbol{s}_j^{(l)}$, can be classified based on its posterior probability. Assuming equal prior probabilities for each cluster, the likelihood function of $\boldsymbol{s}_j^{(l)}$ belonging to cluster *q* is the weighted sum of the likelihoods of $\boldsymbol{s}_j^{(l)}$ belonging to each WTC in cluster *q* in (7-13):

$$p(\boldsymbol{s}_j^{(l)} \mid \omega_q) = \sum_{k=1}^{n(q)} \lambda_k^q \, p_k(\boldsymbol{s}_j^{(l)} \mid \omega_q) \tag{7-13}$$

Where $p_k(\boldsymbol{s}_j^{(l)} \mid \omega_q)$ is a Gaussian function as in (7-3). The posterior probability can be obtained by (7-14):

$$P(\omega_q \mid \boldsymbol{s}_j^{(l)}) = \frac{p(\boldsymbol{s}_j^{(l)} \mid \omega_q)}{\sum_{r=1}^{Q} p(\boldsymbol{s}_j^{(l)} \mid \omega_r)} \tag{7-14}$$

Where $P(\omega_q \mid \boldsymbol{s}_j^{(l)})$ is the posterior probability of sample $\boldsymbol{s}_j^{(l)}$ belonging to cluster *q*.

In summary, the sample load data of a customer will firstly be decomposed as the same procedure. Each daily component will be assessed by posterior probability of each cluster. It will be allocated to the one with highest posterior probability as in (7-15).

$$\hat{\omega}(\boldsymbol{s}_j^{(l)}) = \arg\max_{\omega q}(P(\omega_q \mid \boldsymbol{s}_j^{(l)})) \tag{7-15}$$

Where $\hat{\omega}(\boldsymbol{s}_j^{(l)})$ is the final classification of $\boldsymbol{s}_j^{(l)}$.

In our case, for multiple levels of *(l)*, the classification would be on 5 levels, magnitude, A, D1, D2, and D3: $\hat{\omega}(\boldsymbol{s}_j^{(M)}) - \hat{\omega}(\boldsymbol{s}_j^{(A)}) - \hat{\omega}(\boldsymbol{s}_j^{(D1)}) - \hat{\omega}(\boldsymbol{s}_j^{(D2)}) - \hat{\omega}(\boldsymbol{s}_j^{(D3)})$. If a customer's is classified as 4-2-7-10-3 (60%), it indicates the magnitude is assigned to cluster 4 of magnitude TC, A component is assigned to cluster 2 of A TC

and so on. The probability 60% is calculated as the number of days belonging to 4-2-7-10-3 over the total number of sample days. It indicates the degree of consistency of classification over days.

7.6 Results

This section will firstly demonstrate the load profiling results from MRC. Further, based on proposed classification method, it will be compared with industrial method and time-series clustering. The results show improvement in MRC, which are further explained in different cases.

7.6.1 TCs

All 6369 customers are firstly macro-classified into three groups: residential, small and medium enterprise (SME) and others by the data collections. The method is applied to each group separately. For each group, there are 4 decomposition levels and several clusters as shown in table 7-1.

Table 7-1 Number of clusters of each group and decomposition level

	A	D1	D2	D3
Residential	20	11	8	10
SME	19	8	13	8
Other	15	11	8	9

Figures 7-7 to 7-12 show some of the typical components (TCs) as examples. The TCs are from residential customers at the A level D2 level and D1 level.

Figure 7-7 and Figure 7-8 are two clusters from the A level. Grey lines are some sampled daily components from member customers. The red line is the TC of the cluster, which is derived from the average within the cluster. It is seen that the TC in Figure 7-7 shows a typical working class household with low loading level in the day time and peak in the evening. The TC in Figure 7-8 is, however, has a high loading level consistently through a day. It is probably from a household with full-day occupied.

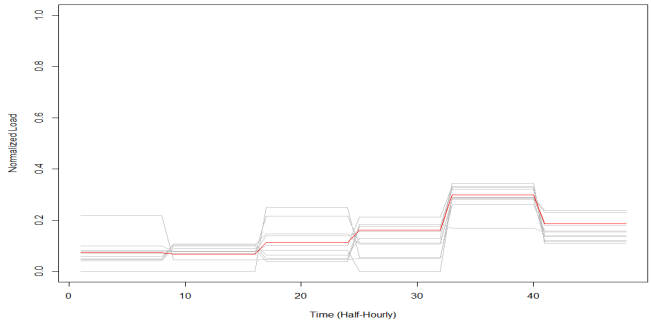

Figure 7-7 TC and members of cluster 1 on A level

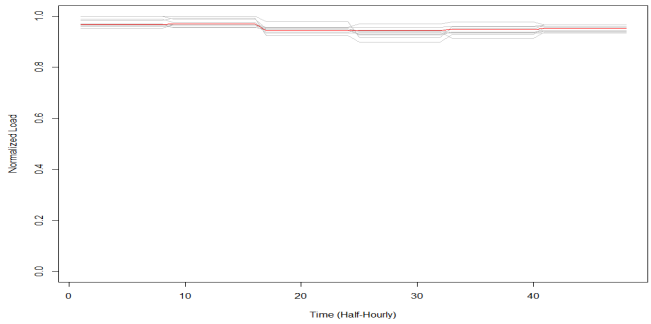

Figure 7-8 TC and members of cluster 2 on A level

Figure 7-9 and Figure 7-10 are two clusters from the D2 level. The scale is getting smaller and there is more variation within the cluster. The components at this level are likely to represent more random activities and short-interval usage (e.g. kettles). Figure 7-9 sees the few activities from 1 a.m. to 8 a.m. (sleeping time) and frequent activities from 9 a.m. The cluster in Figure 7-10 has similar patterns while activities are more concentrated around 7-10 a.m. with a consistent peak.

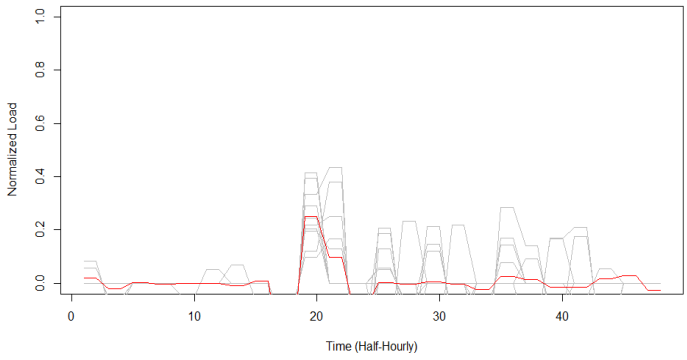

Figure 7-9 TC and members of cluster 1 on D2 level

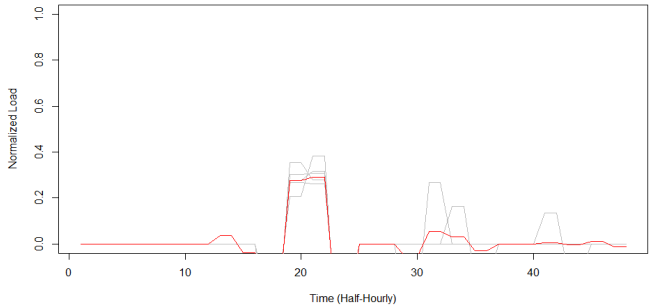

Figure 7-10 TC and members of cluster 2 on D2 level

Figure 7-11 and Figure 7-12 are two clusters from the D1 level, which has the smallest scale. They contain more un-predictable spikes which are possibly caused by the turn-on of some appliances. Cluster in Figure 7-11 has very consistent periodical spikes especially during night and noon, which are probably from stand-by white goods such as refrigerators. In the morning and evening peak times, with human activities, the periodical loads are mixed with other spikes, making the average flatter. Cluster in Figure 7-12 shows no periodical loads, but there are congested high peaks around 10 p.m. till midnight. The small peaks through other time of the day are well dispersed.

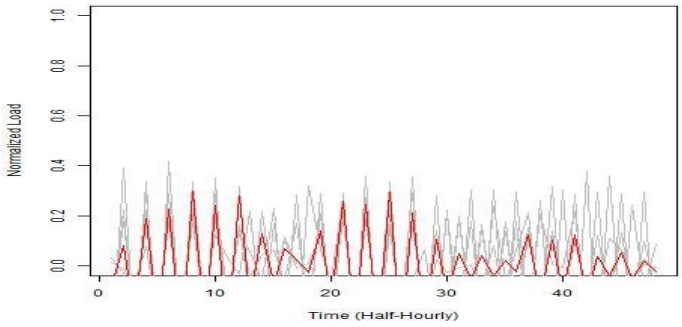

Figure 7-11 TC and members of cluster 1 on D1 level

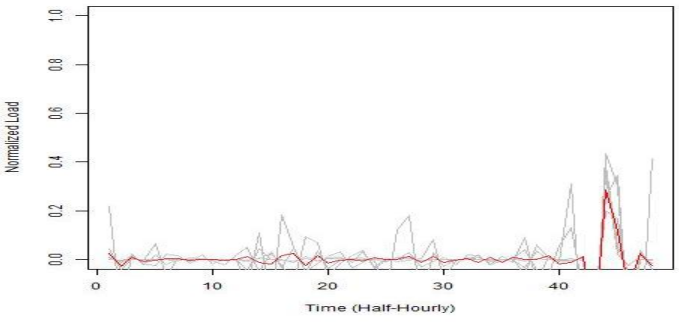

Figure 7-12 TC and members of cluster 2 on D1 level

7.6.2 Comparisons

The comparison is between three types of load profiles: MRC, time-series K-means clustering and TLPs used by the UK industry. It is conducted by assessing the similarities between three load profiles and smart metering data on random days.

In the UK, due to the absence of smart meters on every customer in the market, small customers (below 100 kW maximum demands) are pre-classified into 8 classes for electricity settlement. Each class of customers are represented by a TLP, which has been widely used by the UK industry for decades. The classification is generally based on the nature of customer, such as residential, commercial and industry. Residential and commercial customers are further classified by tariff types while industrial customers are further differentiated by load factors.

The same smart metering data used in the proposed method are also processed by K-means clustering on time domain. The clustering is based on individual load profiles. Each customer each day is assigned with a deterministic cluster and TLP.

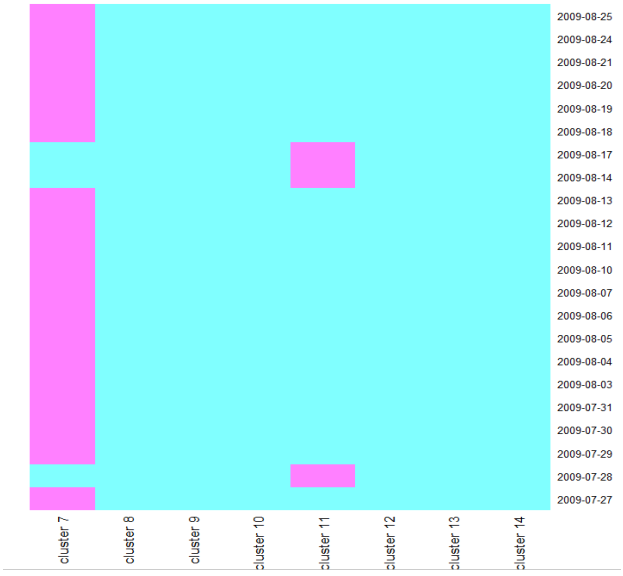

Figure 7-13 Posterior probability spectrum (A level) of customer 1609 over a month

The classification of proposed method is based on a sampled data from sampled days of the investigated customer. Figure 7-13, as an example, depicts the classification spectrum of customer 1609 over a month. The A components of customer 1609 are

classified between cluster 7 and 11 with high probabilities, and thus Figure 7-13 only plots from 7 to 14. It is clearly seen that most (87%) of the days are classified into cluster 7.

Based on the result, the customer is classified as table 7-2. In summary, customer 1609 belongs to i) class 1 in the industry typical load profiling; ii) cluster 6 in time-series K-means clustering (K=30); iii) 7-5-7-6 in MRC (Total number of clusters: 10-5-8-7). We deliberately choose K=30 for K-means, which is the sum of total cluster number in MRC (10+5+8+7), to give a fair comparison between two methods. Because the industry TLPs are represented the average magnitude over millions of customers, and the K-means is based on normalised shape, it is difficult to compare different load profiling methods in terms of magnitudes. The comparison on this chapter is based on normalised load.

Table 7-2 Classification of sampled customer by different load profiling methods

Customer	Date	UK TLP	K-means	MRC (A-D1-D2-D3)
1609	26/Aug/2009	1	6	7-5-6-7

Still, taking customer 1609 as an example, Figure 7-14 shows the comparison between smart metering data and three load profiles. The black line is the real metered data of the customer on the day. The green line is the TLP class 1 from industry. It is unable to express the real daily load profile. The blue line is the cluster 6 in time-series K-means clustering. Although following the base load well, it cannot capture the spike at noon. The proposed MRC, depict in red line, more closely express the daily energy usage in terms of both overall trend and spikes.

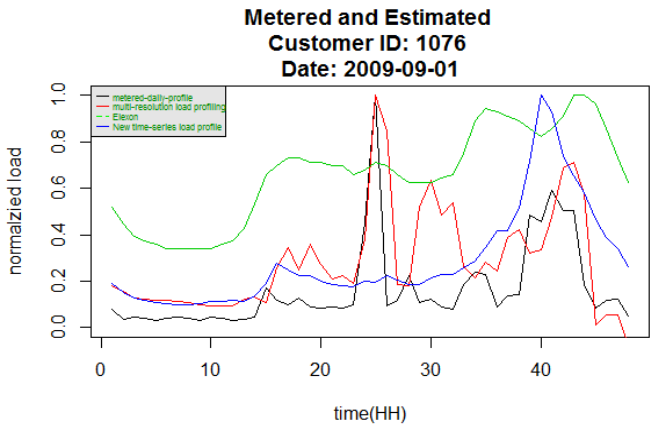

Figure 7-14 Comparison between smart metering data of customer 1609 on 26/08/2009 (black) and three load profiling methods: UK TLP (green), K-means(blue), MRC (red)

The load profiles used by the industry are based on seasonal average load profiles within a group of pre-defined customer class. It is only an approximation of group customers over long term. Moreover, there are only 2 classes describing residential customers, i.e. 2 clusters. For time-series analysis, due to the issues such as volatility and uncertainty, they are more feasible for analysing average load profile over time. The proposed MRC successfully address these issues by separating different load characteristics on different decomposition levels.

An extensive assessment over the three load profiling methods is conducted over 2994 smart-metering load profiles. The representativeness of three load profiling methods are evaluated by the following indices: Maximum Magnitude Error (MME), Mean Absolute Percentage Error (MAPE), Peak Time Error (PTE). The comparison in Table 7-3 shows clear improvements of MRC over the other two methods. PTE has a huge decrease to 2.8 hours. It shows the capability of MRC on capturing the volatilities. MME and MAPE also decrease compared to K-means with similar computations. The reason is that different permutations of sub-TLPs provide a more flexible load profiling with much less computation. By $\sum_{i=1}^{I} n_i$ (30) levels of computation, MRC can express equivalent to $\prod_{i=1}^{I} n_i$ levels load profiling. n_i is the number of TLPs on i^{th} decomposition level.

Table 7-3 Comparison between smart metering load profiles and three load profiling methods (sample size=2994)

Load Profiling Methods	MME (per unit)	MAPE (per unit)	PTE (hour)
UK TLP	0.82	0.45	6.4
K-means (K=30)	0.66	0.23	4.0
MRC(10-5-8-7)	0.58	0.17	2.8

Having an insight into problems with time-series clustering mentioned above, we investigate how the mis-classified load profiles are handled by MRC. Figure 7-15 shows uncertainties between days are resolved by GMM at pre-clustering stage. The left figure shows the non-convex load profiles of the same customer on different days. In MRC, their A components are pre-clustered into two clusters. They are represented

by two WTCs, each as an independent distribution model with all parameters preserved. Also, in Figure 7-16, communication delay issue is resolved by MRC. The A components of two similar load profiles with time-delays are still very similar (right figure). It indicates the overall trend of these two load profiles are the same while small-scale components will probably differ from each other.

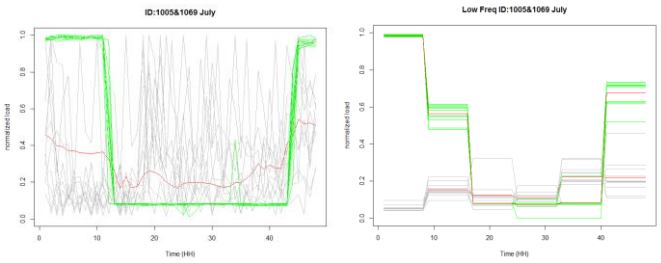

Figure 7-15 Uncertainties between days are resolved by MRC onAlevel

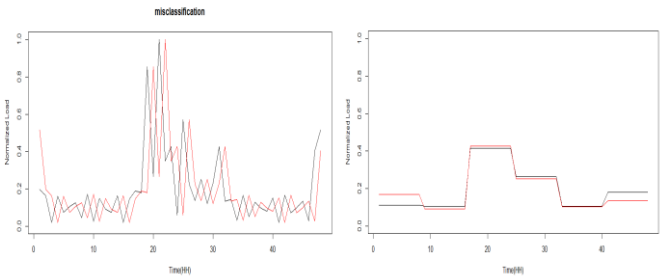

Figure 7-16 Time delay of spikes are resolved by MRC on A level

7.7 Chapter Summary

Based on assessment of spectral analysis techniques on smart metering data in part I, this chapter proposes a novel MRC method. It successfully develops a spectral-domain load profiling for smart metering data. It is specifically designed for smart metering data to overcome the problems caused on traditional time-series analysis.

The result shows significant improvement over traditional time-series analysis, providing more accurate load profiling at more granular level. The proposed MRC offers a promising approach for using smart metering data to develop smart grid and enhance power system efficiency. They can be used for more accurate demand forecast, supporting efficient use of DSR and enhancing settlement (supplier) efficiency.

Chapter 8

Conclusions

THIS chapter draws the conclusion to the thesis by outlining the major contributions and key findings based on the proposed methodology.

The load profiles currently used in the UK for power industry were developed in the 1990s for the purpose of nation-wide electricity market settlement. However, they are unable to accurately reflect: i) the granular energy behaviours for individual customer on individual days; ii) the aggregated load conditions for the LV networks. In order to serve the purpose of LCN and DSR, new load profiling methods are needed to visualise the LV networks and individual customers.

In this thesis, new load profiling methods are developed to contribute to two key areas:

i) For LCN, the traditional indirect load estimation proved to be inaccurate in magnitude at aggregated level and also in shape at more disaggregated level. Yet the direct way, which involves extensive monitoring, can be prohibitively expensive for DNOs. For an accurate but also economical way to visualise the LV networks loading condition, a direct load profiling method is for the first time proposed with three main stages: clustering, classification and scaling. By monitoring and clustering relatively small but representative samples, the common templates are extracted to represent wider areas. A classification tool is also designed to best match un-monitored LV networks to the most similar template without any metering information. In order to reflect both load shape and magnitude, the scaling stage proposes a new peak estimation method, which substantially improves the current estimation accuracy.

ii) It is difficult to directly use smart metering data from individual households to support activities such as DSR, tariff design and load forecast because the data sets are extremely massive, volatile and irregular. Directly applying time-series load profiling would see several limitations in computational burden, accuracy and flexibility. This thesis has proposed another new load profiling method based on spectral analysis, which fundamentally aims to support the big-data analysis. The two main elements in modelling big data were shown to be the number of features and the number of clusters. For feature extraction, excess features will add noise and redundancy into data while insufficient features will lose key information. For classification, a single cluster will mix up everything while too many clusters will lead to misclassification. Instead of

treating them as separate problems, this thesis showed how to find the joint optimal number of features and clusters. Implemented on smart metering data, the method firstly extracts their key features which enable an accurate reconstruction with reduced data size. An innovative MRC technique was proposed to further cluster and classify the extracted features on spectral domain. The proposed method overcomes several limitations with load profiling methods on time-series, and successfully reflects customers' energy behaviours at a granular level.

In summary, the work showed how to extract meaningful information for power system (smart data) from different sources of load information (big data). Due to the different characteristics of the data, two new load profiling methods are developed respectively on time and spectral domain. In detail, the work in this thesis was carried out from four perspectives. The values and limitations are summarised as follows.

LV Substation Clustering and Classification

In order to understand the operation conditions of unmonitored LV substations, LV network templates are developed by using the metered real-time data from selective areas that are representative. By demonstrating on a practical trial UK smart grid project – LV Network Templates, the following observations are reached:

- 10 network load profile clusters with different load shapes are produced by using the metered data. The variances of load shapes are below 0.2 standard deviations within clusters. The distinct characteristics, such as customer composition, geographical information, and customer electricity use behaviour, can be clearly observed in the developed clusters.

- A classification tool is developed to assign un-monitored substations to the appropriate clusters by only using fixed data. The achieved accuracy is 82.2%, which statistically indicates the load shapes of more than 82.2% substations in the demonstration can be predicted with errors less than 0.2 unit standard deviation.

It is the first attempt to visualise LV distribution networks by directly clustering and classifying LV substations. The metering data are massive compared to previous

studies and the classification is difficult as only fixed data are available. The contributions on techniques are:

- Clustering: K-means clustering provides a quick assessment of the optimum number of clusters due to its computation simplicity. However, the results could change with different initial points in K-means. The issue is addressed by hierarchical clustering which provides a deterministic result for given number of clusters.

- Classification: Compared to common pattern recognition techniques, our proposed MLR classification entirely rely on routinely available fixed data, such as customer numbers, feeder length, etc. It does not require any sampled data (metered real-time data) to classify a given substation to a substation template. It is found that the highly mixed LV substations can be distinguished with high accuracy without wide-scale monitoring.

LV Substation Peak Load Estimation

Peak estimation is introduced at the stage of scaling, based on which the load profiling method can provide not only shape information but also accurate magnitude information. The challenge in doing this is that the peak estimation has to solely rely on readily available fixed data so that it is applicable to unmonitored substations. This work proposes an effective CWCR method to estimate the peak demand for LV substations only based on available fixed data. It develops a contribution factor to facilitate cluster-specified peak estimation. The extensive demonstration illustrates that:

- Compared with traditional peak estimation methods, the accuracy and stability of peak estimation has been substantially improved in terms of both R squared error and performance of cross validation.

- As the templates developed by clustering and classification can provide load shape information of LV substations, the peak estimation/scaling can inform their magnitudes. Rather than providing annual peaks, the proposed method is effective in estimating daily peaks.

- The LV templates have significantly reduced massive LV networks into representative models where all kinds of analysis can be efficiently conducted. The work can enhance the understanding of conditions of unmonitored LV substations. It is particularly useful into the future to understand their capabilities to accommodate the increasing penetration of LCTs.

The contribution from developed techniques are summarised as contribution factor and CWCR:

- A contribution factor is proposed to reflect the contribution from a particular customer to the peak of different types of LV substations. As a LV substation usually serves different type of customers, not every customer's peak load coincides with the aggregated substation peak. Customers contribute differently to LV substation peak. Moreover, even the same customer can contribute differently to different types of LV substations. The contribution factor is firstly brought up to describe the diversifications of both customers and substations.

- CWCR is developed to properly separate data and conduct estimation by considering practical situations: i) the intercept should be zero as there would be no load if customer number is zero; ii) all customer classes should contribute positive load to the substation load and therefore the coefficient should be non-negative; iii) a weight is assigned to reflect the variances within clusters.

Assessment of Spectral Analysis Techniques on Feature Extraction and Data Compression

The spectral analysis provides an opportunity to analyses load profiles in the spectral domain. The direct advantage includes data reduction and feature extraction. A comprehensive review and assessment on two spectral analysis techniques including two extreme cases have been presented. In DWT, *Haar* is chosen as the most compact wavelet while sinusoidal wave in DFT gives a global support. A new data reduction and load characterisation method based on DWT is proposed for load profiles of smart metering customers. The key findings are:

- for smart meters, DWT decompose their daily load profiles into more meaningful

and consistent components compared with DFT;

- for smart meters, DWT was found to be more effective and reliable for data reduction than DFT. DWT can reconstruct the original daily load profiles using less data while maintaining high representativeness;

- for smart grid data, where the load profiles are aggregated over time or population, the performance of DFT can be substantially improved with the increase of aggregation level. However, customer size is less influential to DWT than DFT based on the assessment;

- the performance of DFT was shown to become stable and superior to DWT when the aggregation level is sufficiently high.

The contribution from developed techniques are summarised as data compression and feature extraction through DFT and DWT:

- Formalise the major steps for compression and feature extraction of load profiles by DFT. As the frequency increases, the magnitude of component dramatically drops. Aggregation of the first few DFT components is expected to capture the original load profile with high accuracy while the data size can be significantly reduced. The work fixed a flaw found in previous research about the Nyquist frequency, which was considered as part of DC, but is actually a triangular wave.

- Formalise the major steps for compression and feature extraction of load profiles by DWT. DWT coefficients, especially the small-scale ones, have very low magnitudes through most of the time windows. Thus, coefficients below a pre-defined threshold will be set as zeros. By this way, the number of non-zero coefficients can be significantly reduced.

Multi-resolution Clustering

The time-series clustering shows limitations in handling big data, volatility and uncertainty. This thesis proposed a novel MRC method. It successfully developed a spectral-domain load profiling for smart metering data, which overcomes the problems caused on time-series analysis:

- Multi-resolution analysis based on wavelet analysis decomposes load into different characteristics and runs clustering analysis separately. It addressing the volatility in smart metering data.

- A two-stage MRC technique is proposed to cope with uncertainty and reduce clustering input size.

- Different permutations of sub-TLPs provide a more flexible load profiling with much less computation.

The result shows significant improvement over traditional time-series analysis, providing more accurate load profiling at more granular level. The contribution from developed techniques are summarised as:

- Multi-resolution analysis decomposes volatile and irregular load profile into a smooth large scale component describing the underlying shape, and several small scale components describing volatilities. Each clustering will focus on one characteristic without interference between each other. The input data size is also substantially reduced and it provides possibility for different number of clusters on different level.

- GMM is used to pre-cluster each customer through multiple days so that one customer's daily profiles are represented by several typical Gaussian models. Then X-means clustering is performed only on these typical models. By the two-stage clustering, massive inputs are reduced into a small number of PDFs while original information is maintained in the models.

- As the GMM saves full information of each model, the new clusters after second-stage are still PDFs with new parameters, which can be used to calculate the posterior probability of a new sample. It provides a probabilistic classification for new smart metering customers.

Chapter 9

Future Works

THIS chapter presents future work that can be done to improve the investigations of smart tariff designs and applications.

Applications of LV Network Templates

The work has developed the LV network templates to visualise the LV network utilisation. Based on the classification of LV networks and the estimation of load profiles, they can be widely used as an effective tool and platform for different analysis in power system:

i) To develop network specified DSR by commercial and technical solutions: LV networks have a large variety of customer mix, with very different load shapes and network stress points. Therefore, the same individual response will reflect different effects on different LV networks. For each template, we will evaluate the performance of different intervention combinations on addressing the network stresses. A unique commercial solution (e.g. tariff) should be developed for each type of networks. For technical solutions, such as network reconfiguration and interconnection, it is also critical to match the most complementary pair of substation types.

ii) To develop network specified Use-of-System (UoS) charge. The current network charges only consider the investment based on annual peak, but cannot reflect the real-time loading condition of the network. In order to guide customers to use the spare capacity in the trough time and reduce load during the peak, it is critical to develop variable UoS charges which reflect the typical loading condition of the network. Such network specified UoS will not only give locational signals for investment, but also temporal signals for operation.

Customer classification based on behaviour modes and social-economic status

By managing individual customer's energy usage, the aggregated effects can significantly enhance the power system efficiencies and reduce the carbon emissions. However, the diversity of customers' load profiles makes it difficult to identify their behavior modes and response potentials. In fact, load profiles directly reflect customers' energy behaviors (habits), which are potentially rooted in their social-economic conditions. To better understand this chain and thus manage customers' energy usage, future work will investigate the relationship between energy behavior

and social-eco status. The current researches are mainly separated between energy patterns, behaviours and socio-economic information.

The rich data from smart meters are now providing an insight into the customers' energy usage characteristics. However, recent studies mostly focus on decomposing load into different electricity appliances but not into different energy behaviours. Such detailed appliances decomposition can only provide retroactive records of customer's energy usage while most customer energy management requires proactive signals and steps. For example, when new customer classifications are informed by samples of smart metering data, there is the need to i) allocate a customer to relevant class, ii) to forecast his energy usage, iii) to identify his energy behaviour entirely based on customers' social-economic characteristics. Thus three steps of work will be conducted in the future:

i) Decompose load into different components, where each can be linked with certain customer behaviours with different confidence

ii) Match the typical behaviours with customers' social-economic status as they to some extent represent the lifestyle and social class of a household;

iii) Develop an energy classification. Similar with social class, customers can be allocated to certain energy behaviour types based on their social-eco information.

Whole system analysis

In DSR, it is critical to consider the balance of benefits for the whole system. For power system in particular, the balance can be seen in two main streams:

i) The balance between different voltage levels: the change of customers' energy usage will simultaneously affect supplier, generation and networks at different voltage levels. However, the different parts of power system hardly share the same interest simultaneously. For example, the coincidence of a LV network to its upper stream HV network can be very low. Thus when a customer responds the HV network, he could possibly shift the load to the pressure point of the LV network.

ii) The balance between different signals: DSR can have multiple objectives such as reducing energy price, releasing network pressure, regulating reactive power and controlling frequency. Again, when customer responds to a particular signal, it could unintentionally exert a negative influence to another objective.

Based on the customer and LV network classification above, I could conduct a whole system analysis, aiming to reduce massive combinations between customer mixes and network models down to a manageable number of typical models. Each model would share similar: network pressure (magnitude/time), customer mix, customer response, and thus management strategy. We will also investigate the conflicts between network pressure, energy price and customer behaviour. By cluster-specific business model and shared-control technical solution, an optimal strategy could be developed to reduce network pressure meanwhile ensuring customers' benefits.

Big Data analysis to interconnect different datasets

With the development of LCT and smart meters, big data from other industry sectors will become available. More importantly, they will be increasingly linked with the energy section with the development of technologies such as EVs. Therefore, the whole system analysis can be extended to different energy sectors, which essentially means to discover the underlying relationships between data sets from different sources.

The socio-eco information can thus act as the core data which connect with data from all types of industries. Firstly, from the smart metering data, customer behavior can be inferred, which can be further used to extract the socio-eco information. Similarly, the socio-eco information can be extended to explain human's behavior in other industries. It actually acts as a base station to interconnect different datasets.

Appendix. A

Data Sense-Checking

Before going through date clustering and interpretation, three main issues were detected with the data received: i) low resolution, low currents are represented only by a few limited readings, ii) there may be problems with these low current readings, iii) the power calculated from voltage and current readings and 'real power delivered' conform well at most high current substations, but there are still a number of substations where the two powers differ significantly; and iv) at some substations the majority of power readings are zero.

There are some issues with current readings, particularly with current readings that are low. Figures A-1 and A-2 show the occurrence frequency of different current readings at all substations with different ranges. Three main issues are observed.

Poor resolution at lower currents. The readings are more discrete at low currents. As seen from Figure A-1, the readings of 8.763561A and 10.73313A appear approximately 200,000 and 100,000 times respectively while there are no other readings between the two values. This situation continues on readings as high as 20A.

Different resolution levels throughout the metering range. It is demonstrated in Figure A-2 that as the current grows, the resolution increases as well. At the lower end, the resolution is around 2A, which decrease to 1A then 0.5A as the current increases. The resolution reaches its minimum of approximately 0.2A when the current is between 90A and 500A, but, it increase back to 0.4A again when the current is higher than 600A.

Regular frequency peaks. It is also obvious in Figure A-2 that there are some current readings with much higher occurrence frequency compared with the adjacent readings. These high-occurrence currents are normally peaks, and the interval values between two peaks decreases as the current readings increase. Besides, between two peaks, there are some sub-peaks dominating a smaller time interval. This trend is consistent throughout the range of the whole metered data.

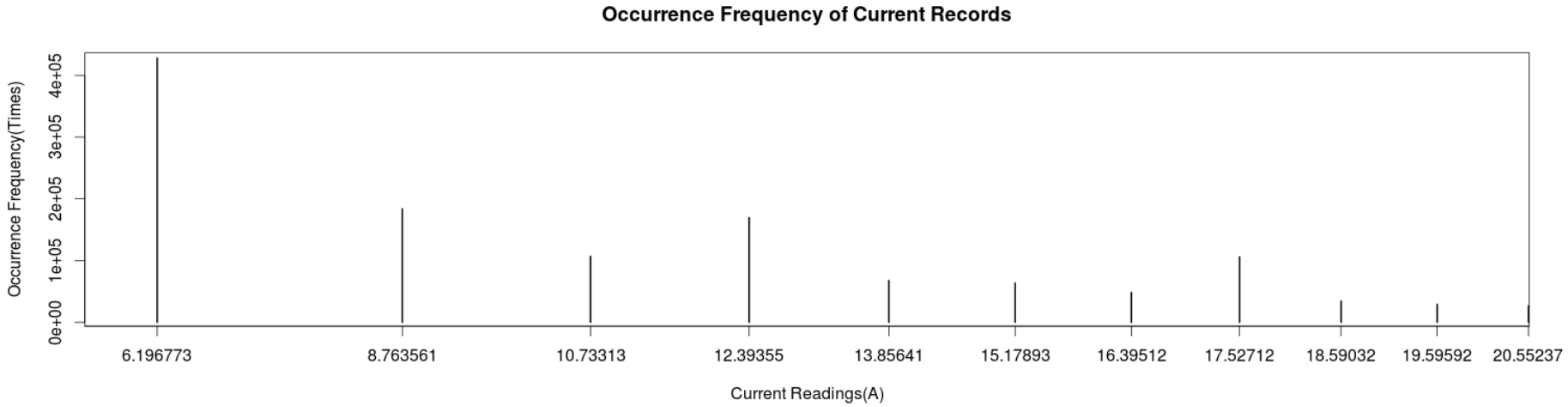

Figure A-1. The Occurrence Frequency of Current Records from 0A to 20A

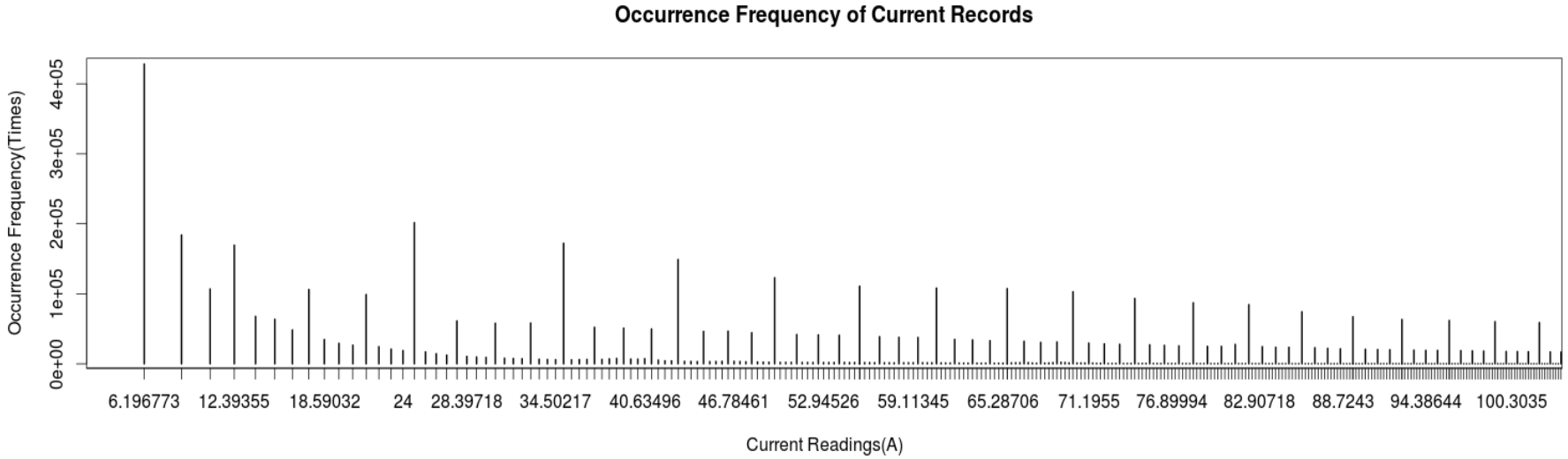

Figure A-2. The Occurrence Frequency of Current Records from 0A to 100A

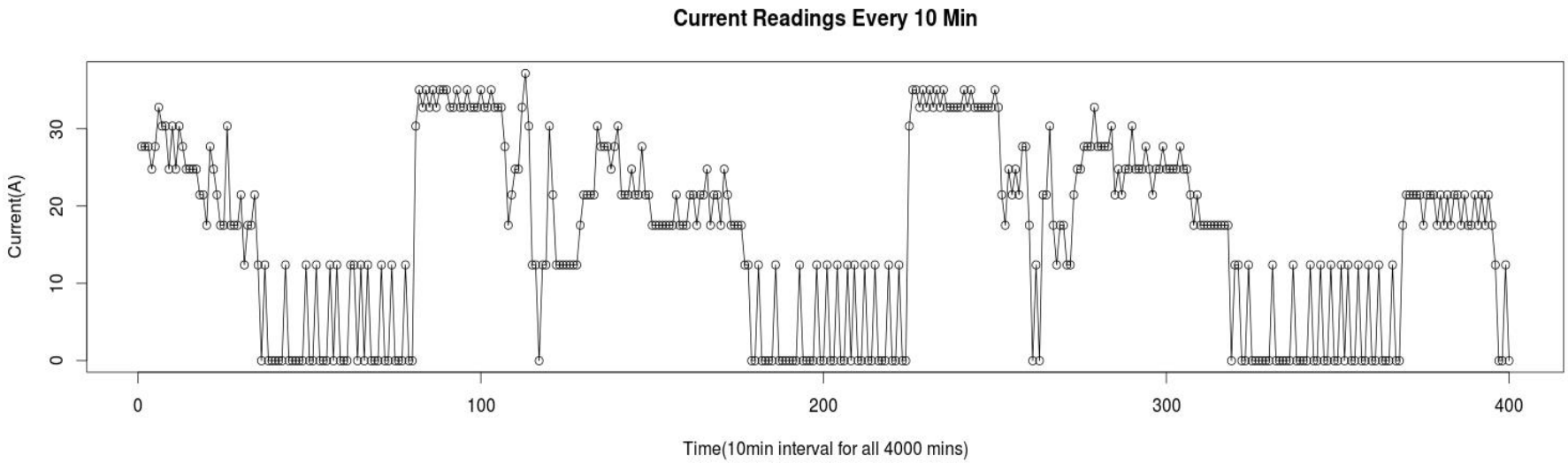

Figure A-3. The Every 10 min Current Records for Substation 536787

Theoretically, the real power delivered is equal to the power calculated from measured voltages and currents multiplied by a power factor. The value of power factor ranges from 0 to1 but usually higher than 0.9, which means the metered real power delivered should be very close to the real power calculated. However, the comparison of the two power at all substations indicates diversifications. Roughly, all substations can be categorized into three groups according the difference between the real power calculated and delivered.

Conforming group

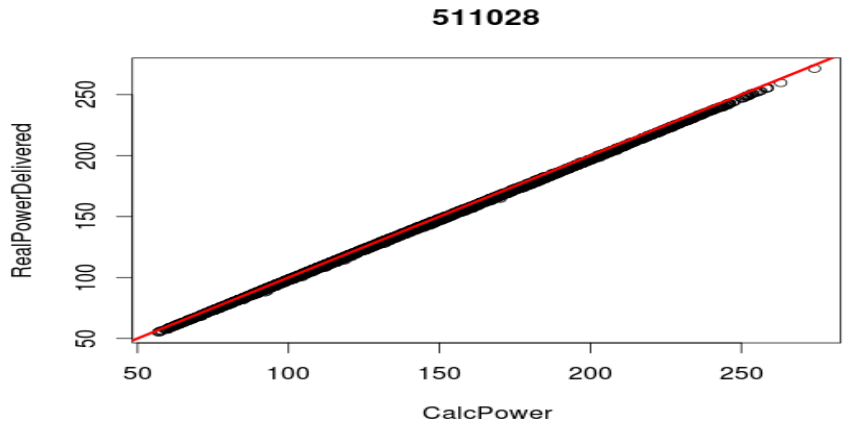

Figure A-4. Power Comparison Scenario 1: conforming group

Low current and power group

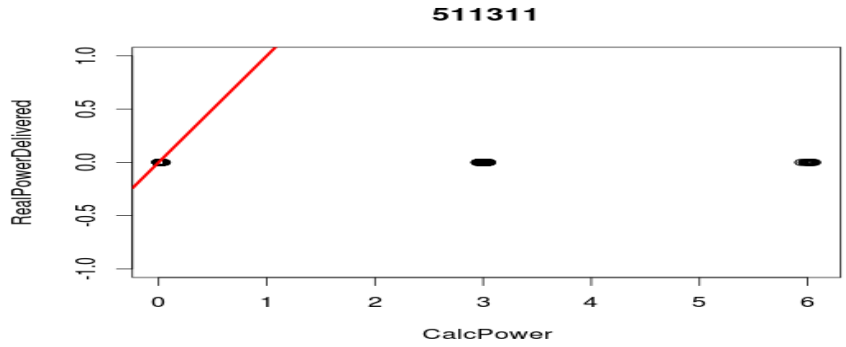

Figure A-5. Power Comparison Scenario 2: zero current group

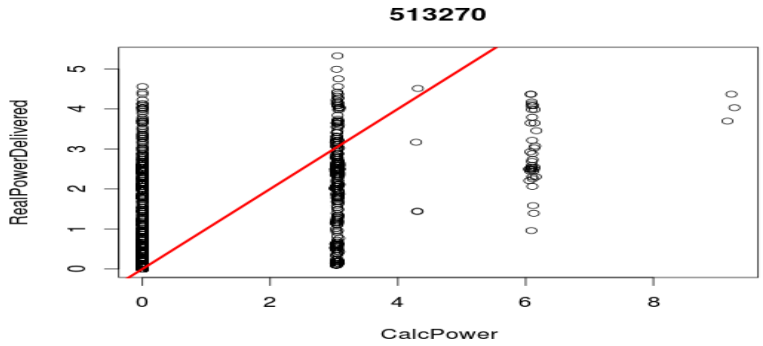

Figure A-6. Power Comparison Scenario 2: discrete power

Suspicious group

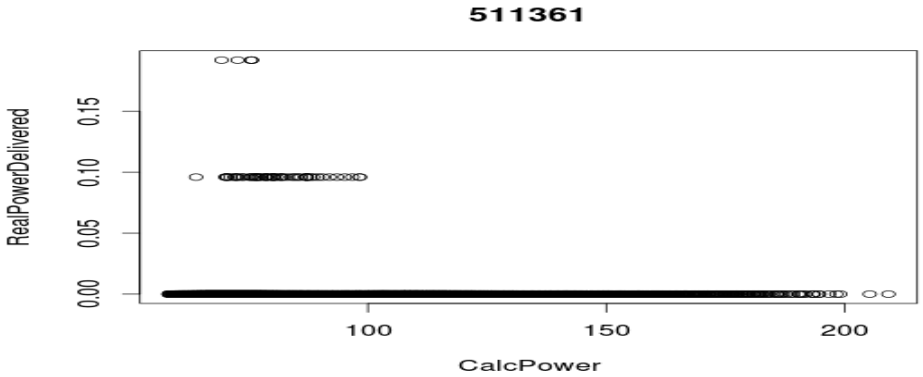

Figure A-7. Power Comparison Scenario 3: one zero power readings

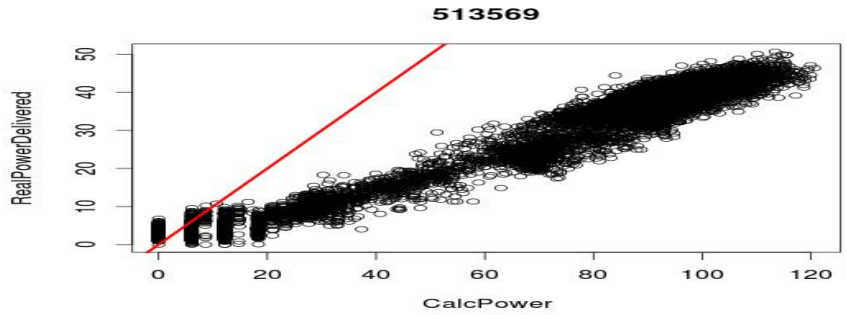

Figure A-8. Power Comparison Scenario 3: Wrong Ratio

The very low demand subs were supplying spurious readings as the current was so low that the CT ratio used was too large for the meter to accurately record usage. A current amplifier was designed by Haysys and fitted on 17 of the worst substations. This was successful and the resulting readings were divided by 10 to give the correct current.

Appendix. B

The parameters are estimated by maximum likelihood Estimation (MLE). The likelihood function is defined as

$$L(b) = \prod_{i=1}^{I} P(Y_i \mid b) \tag{B-1}$$

where, $b = (b_1 ... b_K)$ is the matrix of coefficients to be estimated for all K clusters; $L(b)$ is the likelihood function; I is the total number of observations (substations); $P(Y_i \mid b)$ is defined in (B-2)

$$P(Y_i \mid b) = \prod_{n=1}^{K} P(Y_i = n)^{g(Y_i = n)} \tag{B-2}$$

where, $g(Y_i = n)$ is the indicator function which is equal to 1 if substation Y_i belongs to cluster n, or 0 otherwise.

In practice, it is more convenient to use log-likelihood in (B-3) derived from (B-2), which is globally concave [102].

$$\ln L(b) = \sum_{i=1}^{I} \sum_{n=1}^{K} g(Yi = n) \cdot \ln P(Y_i = n) \tag{B-3}$$

Making use of unity probability, $\sum_{n=1}^{K} g(Y_i = n) = 1$ with (4-10) and (4-11), the log-likelihood function can be written in (B-3).

$$\ln L(b) = \sum_{i=1}^{I} [-\ln(1 + \sum_{k=1}^{K-1} e^{b_k \times z_i}) + \sum_{n=1}^{K-1} g(Y_i = n) \cdot b_n \times z_i] \tag{B-4}$$

The objective is to find the coefficient b in (B-4), which maximises the log-likelihood function. The optimization problem can be solved by Newton methods which find the stationary point of the gradient of the log-likelihood. The detailed algorithm can be found in [103] and it is proved to have a guaranteed convergence for MLE [104].

In MLR, a linear predictor function is used to relate variables to logistic index V_{ik}.

$$V_{ik} = b_k \times z_i = b_{0k} + b_{1k} \times z_{i1} + b_{2k} \times z_{i2} + ... + b_{mk} \times z_{im} \quad \text{(B-5)}$$

where, $z_i = (z_{i,1}...z_{i,m})$ is the i^{th} set of m independent variables; $b_k = (b_{0,k}...b_{m,k})$ is the regression coefficient for cluster k; and V_{ik} is the logistic index of observation i belonging to cluster k.

Taking cluster K as reference (baseline) and V_{iK}=0, $P(Y_i = n)$ is the probability that the i^{th} observation belongs to cluster n, then

$$P(Y_i = n) = P(Y_i = K) \cdot e^{V_{in}}, \qquad n = 1,2,...K \quad \text{(B-6)}$$

Consider the fact that the probability of an observation i belonging to all K clusters must be 1 shown in (B-7).

$$\sum_{n=1}^{K} P(Y_i = n) = \sum_{n=1}^{K} P(Y_i = K) \cdot e^{V_{in}} = 1 \quad \text{(B-7)}$$

By restructuring (B-7), (4-10) and (4-11) can be easily derived.

Appendix. C

LV Network Templates in Summer and Autumn

Template 1

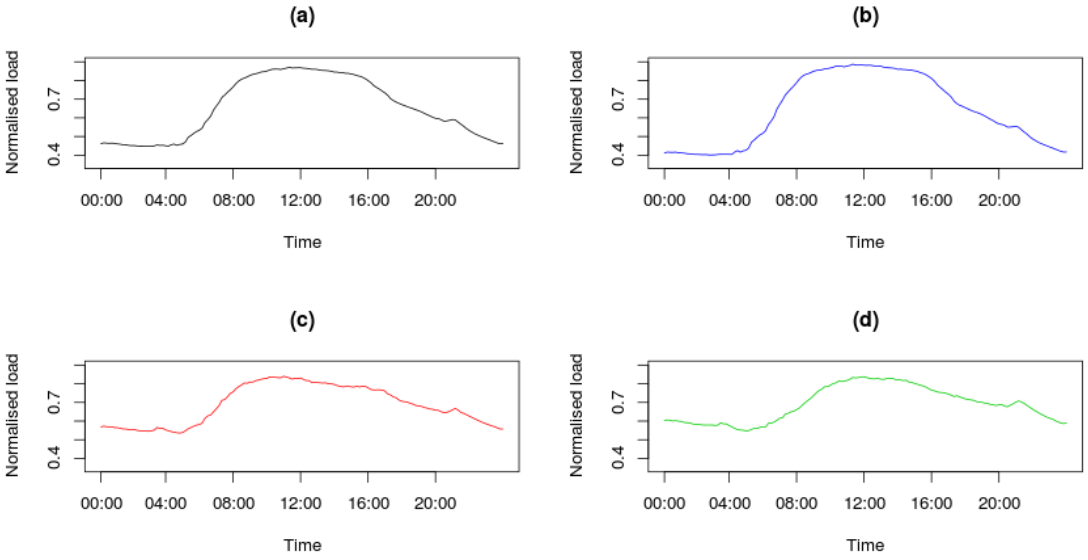

Figure C-1: Substation demand profiles for cluster 1 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

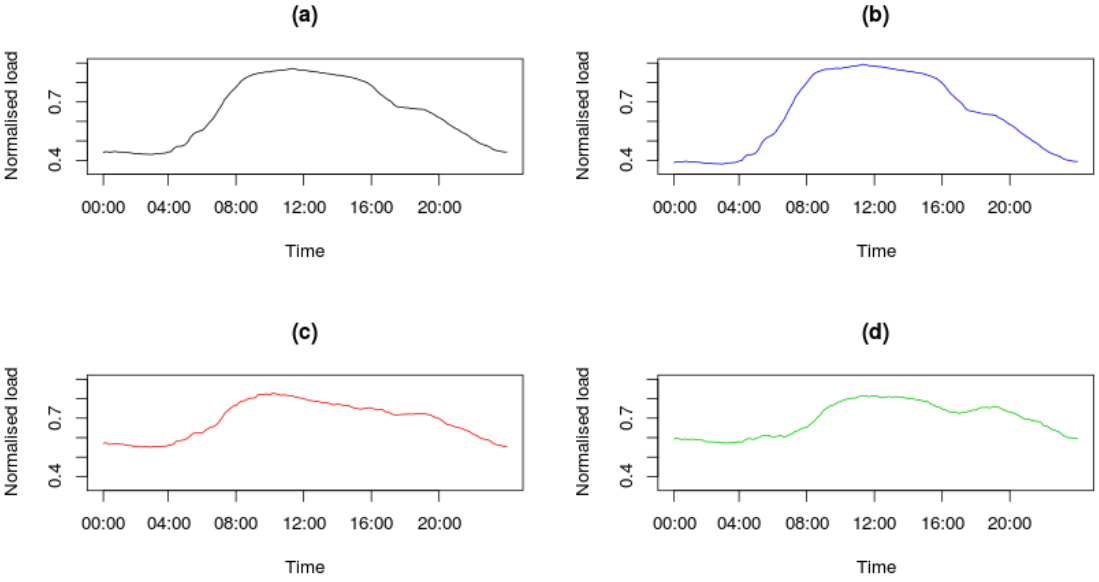

Figure C-2: Substation demand profiles for cluster 1 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 2

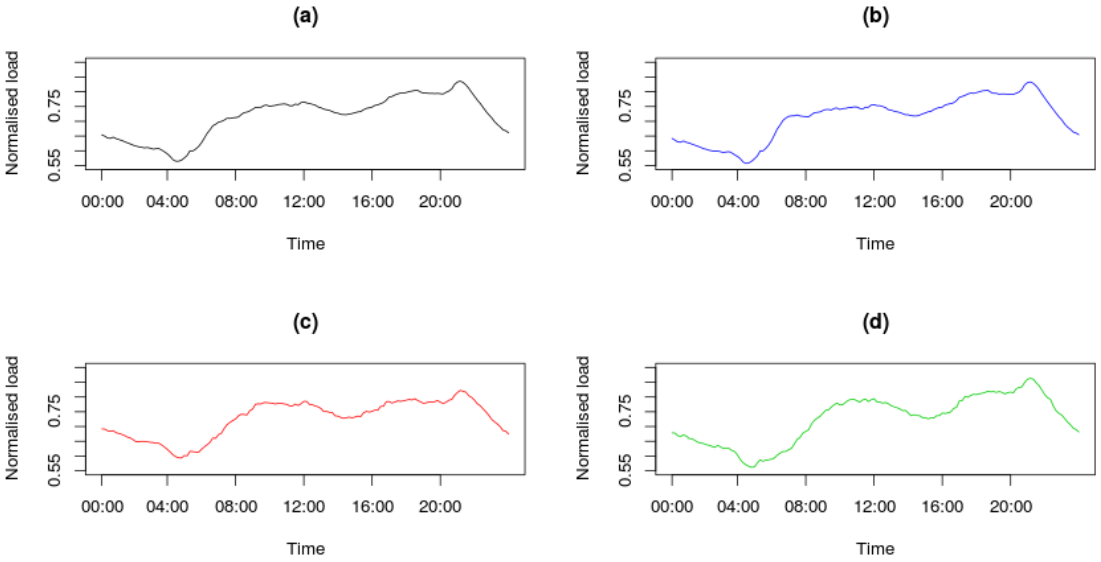

Figure C-3: Substation demand profiles for cluster 2 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

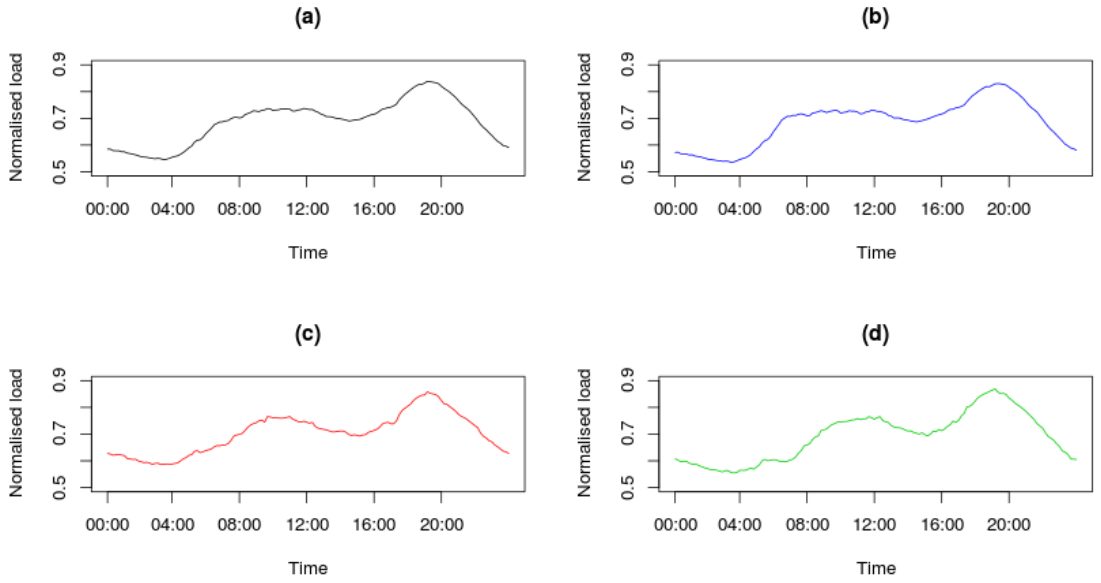

Figure C-4: Substation demand profiles for cluster 2 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 3

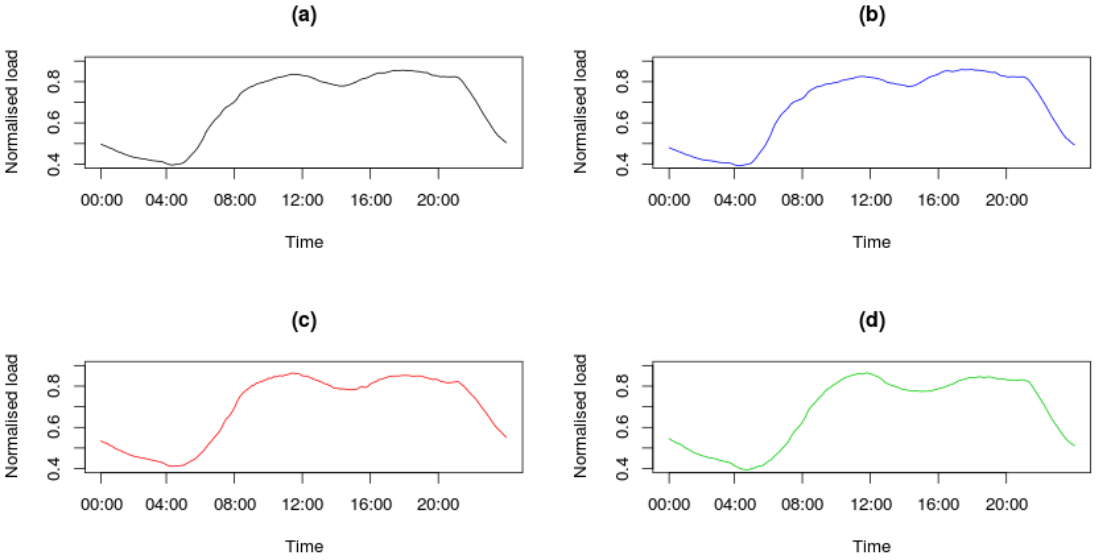

Figure C-5: Substation demand profiles for cluster 3 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

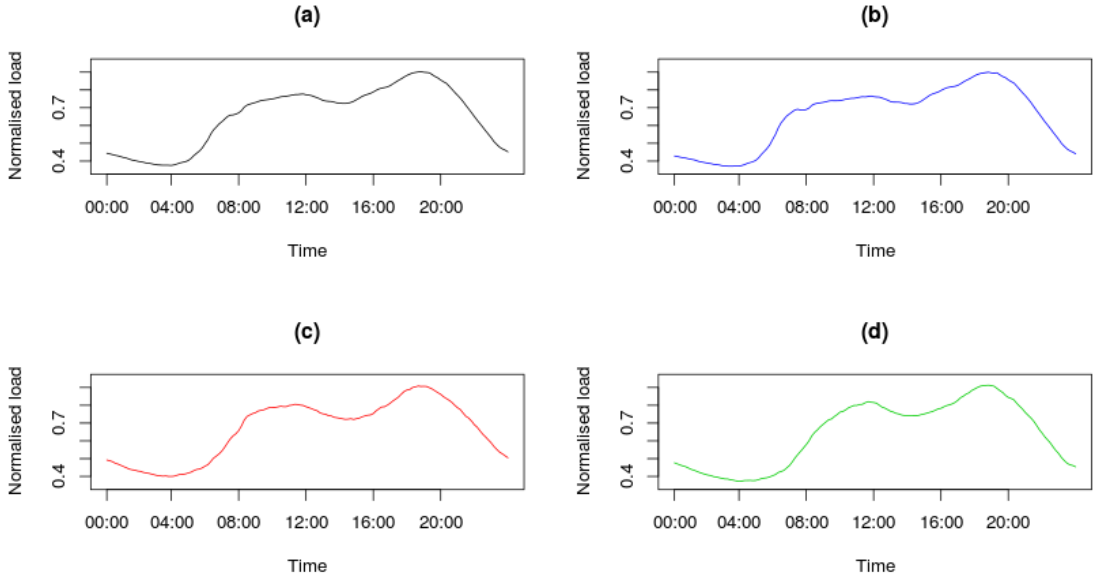

Figure C-6: Substation demand profiles for cluster 3 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 4

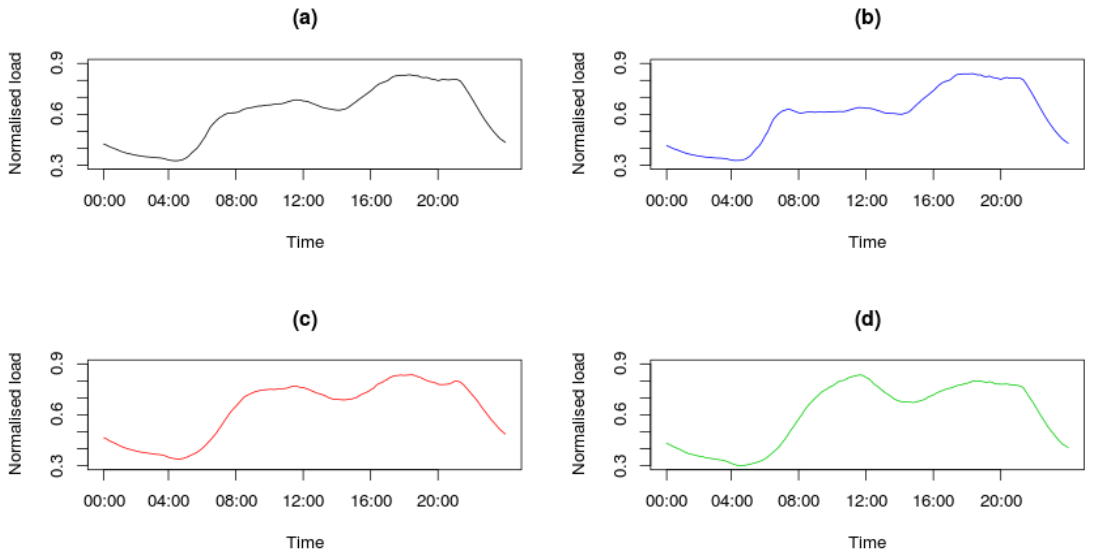

Figure C-7: Substation demand profiles for cluster 4 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

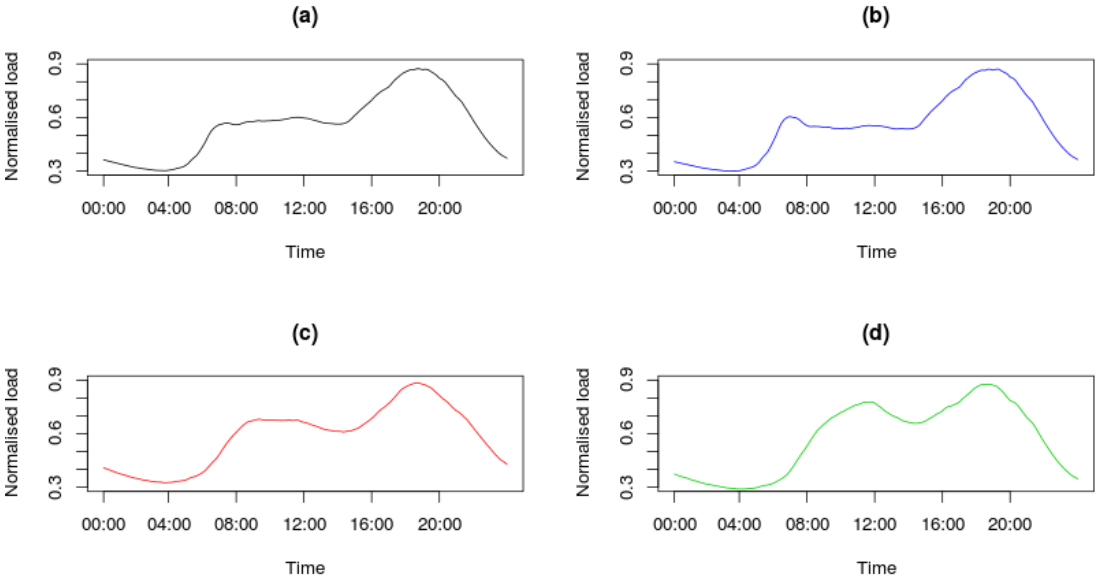

Figure C-8: Substation demand profiles for cluster 4 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 5

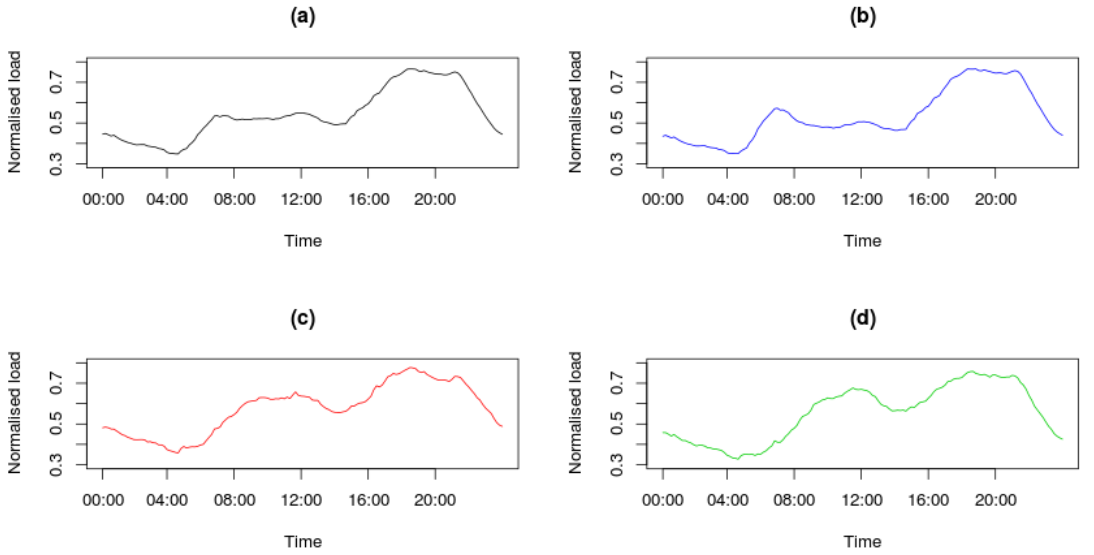

Figure C-9: Substation demand profiles for cluster 5 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

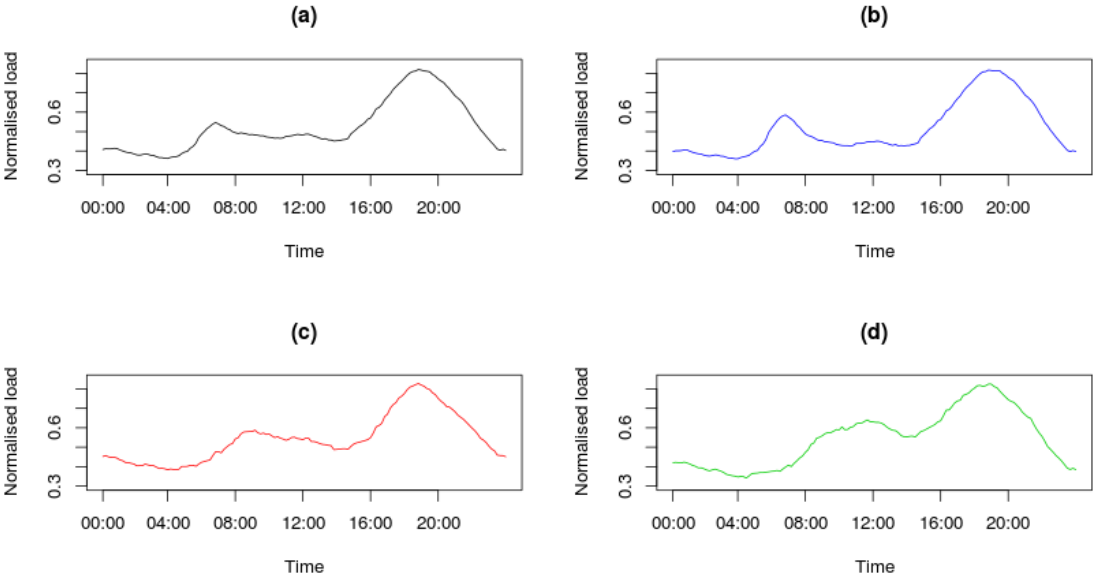

Figure C-10: Substation demand profiles for cluster 5 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 6

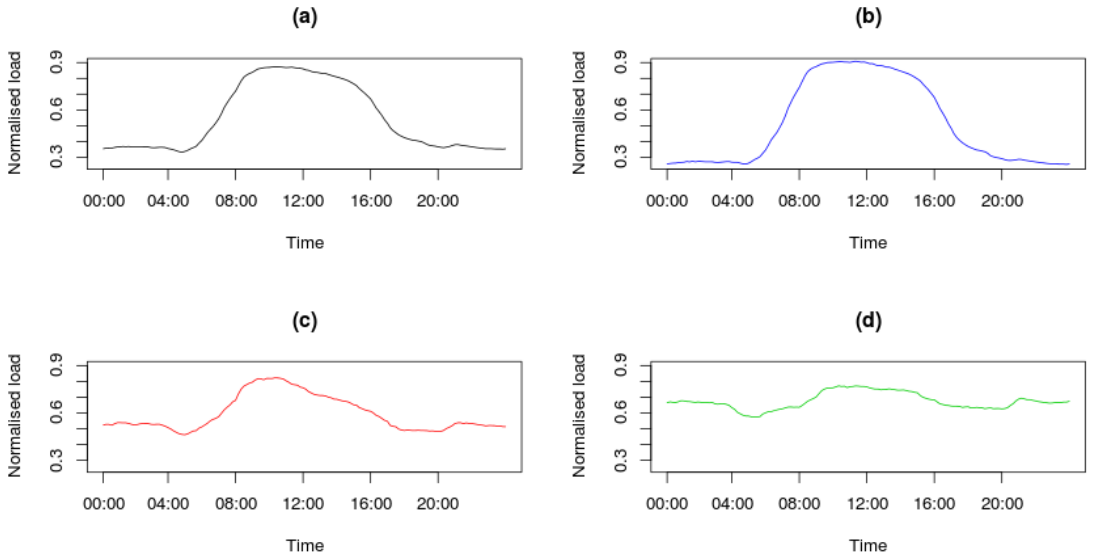

Figure C-11: Substation demand profiles for cluster 6 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

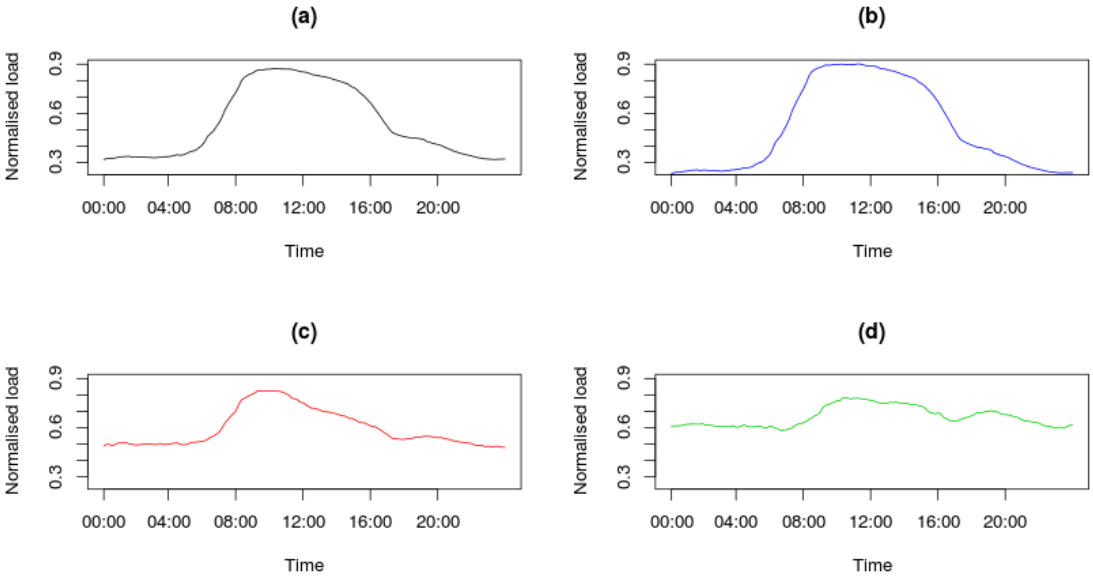

Figure C-12: Substation demand profiles for cluster 6 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 7

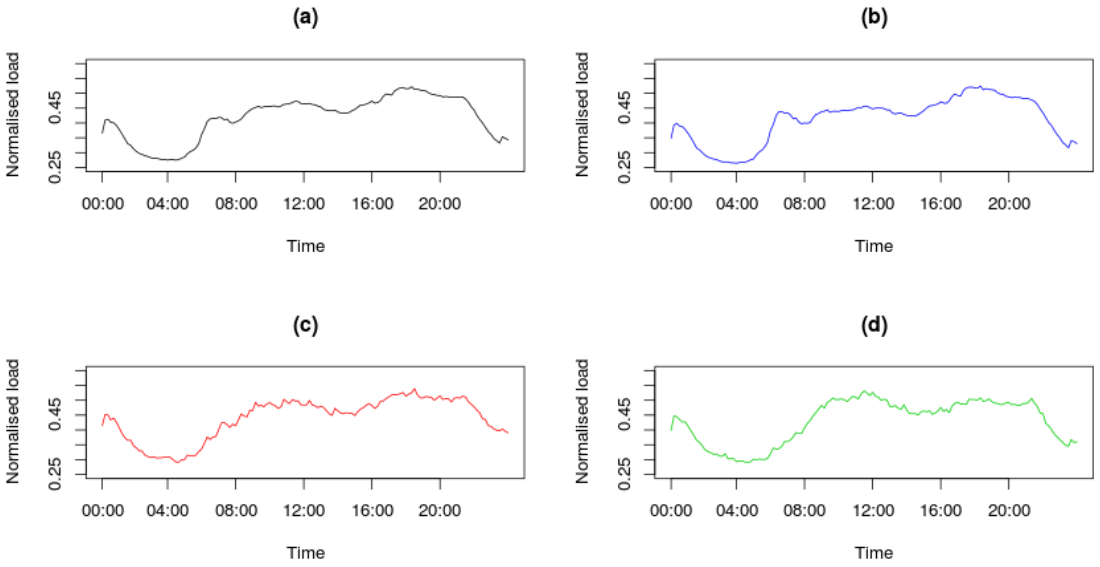

Figure C-13: Substation demand profiles for cluster 7 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

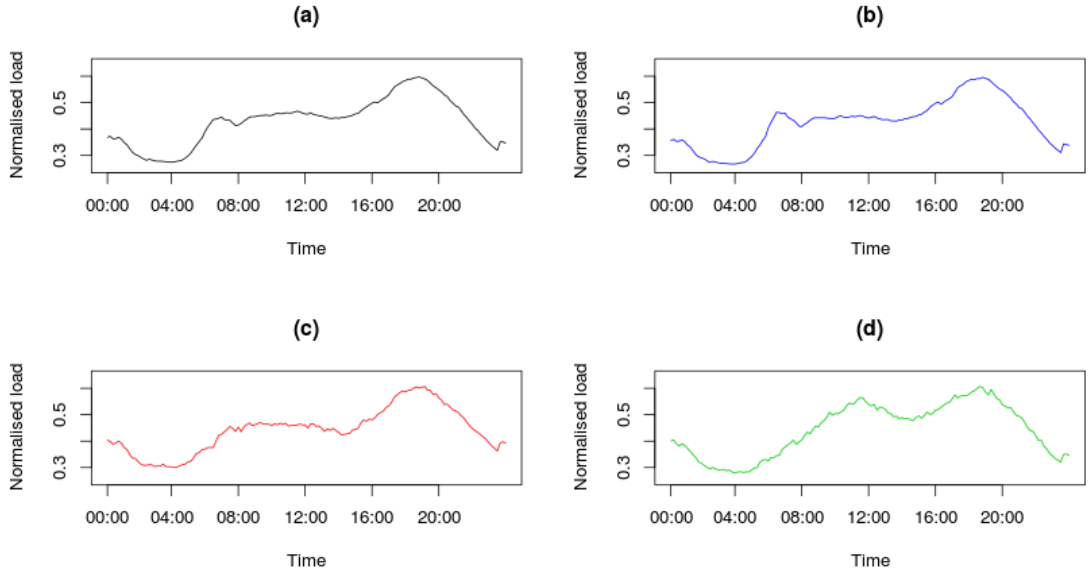

Figure C-14: Substation demand profiles for cluster 7 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 8

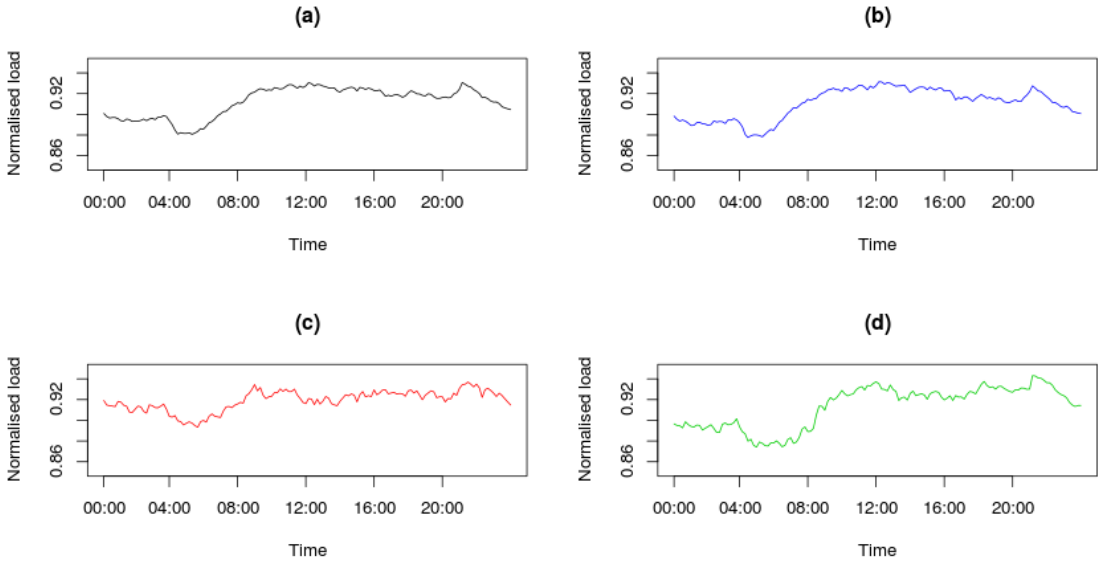

Figure C-15: Substation demand profiles for cluster 8 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

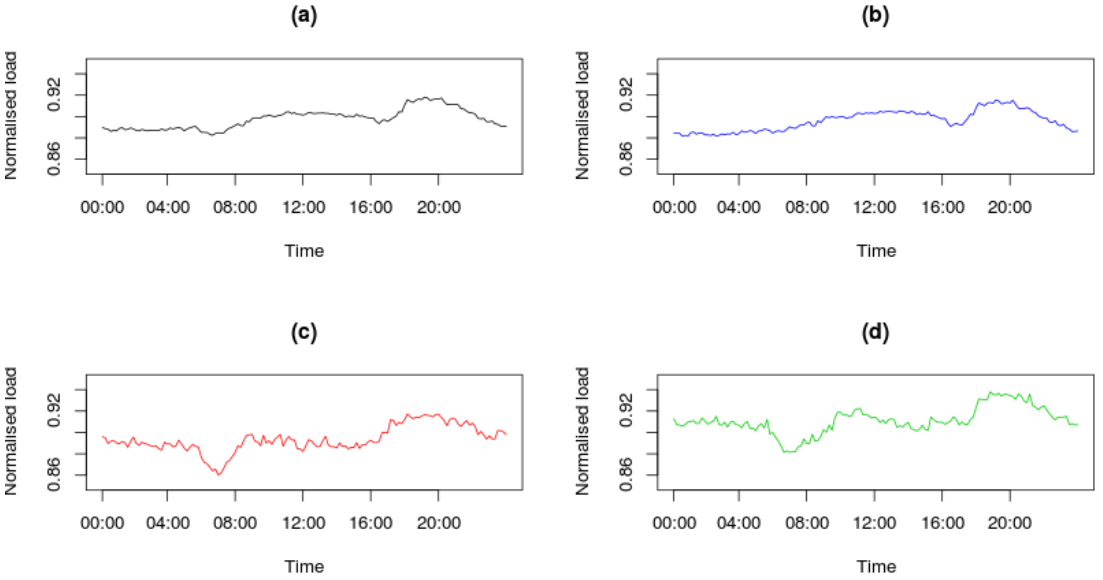

Figure C-16: Substation demand profiles for cluster 8 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 9

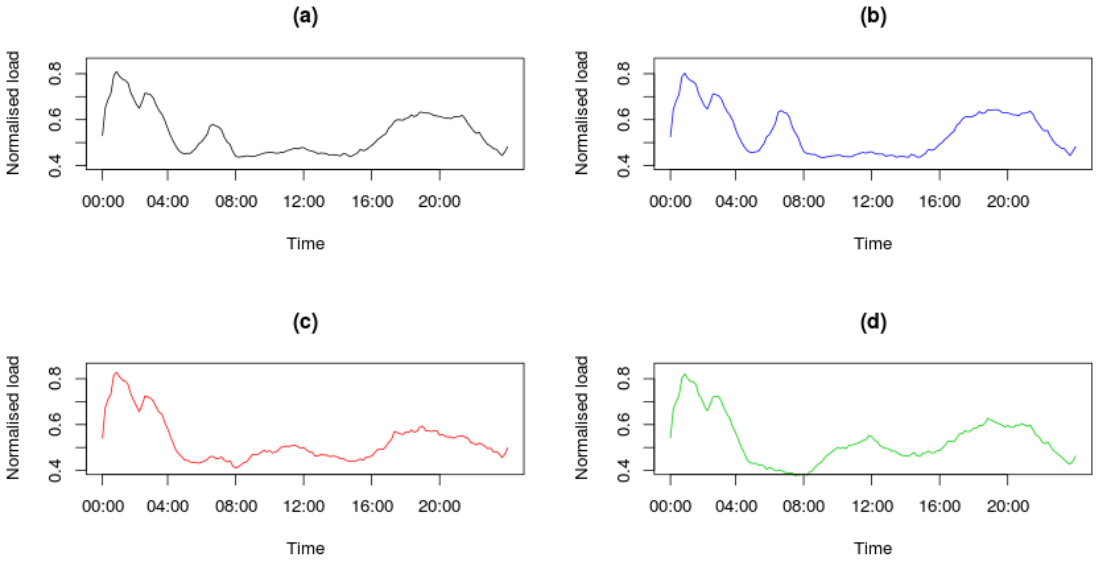

Figure C-17: Substation demand profiles for cluster 9 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

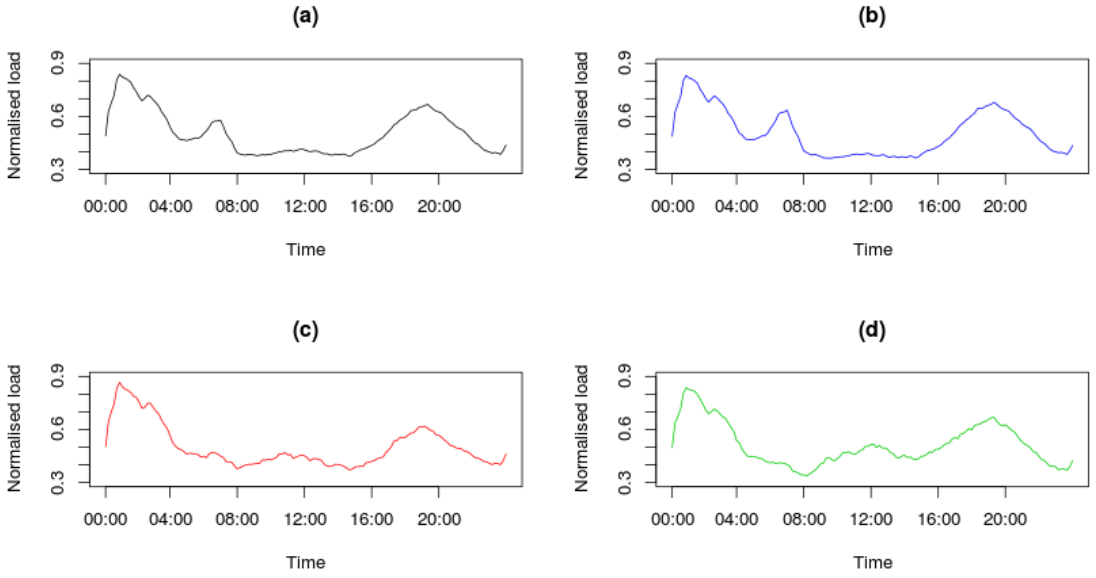

Figure C-18: Substation demand profiles for cluster 9 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Template 10

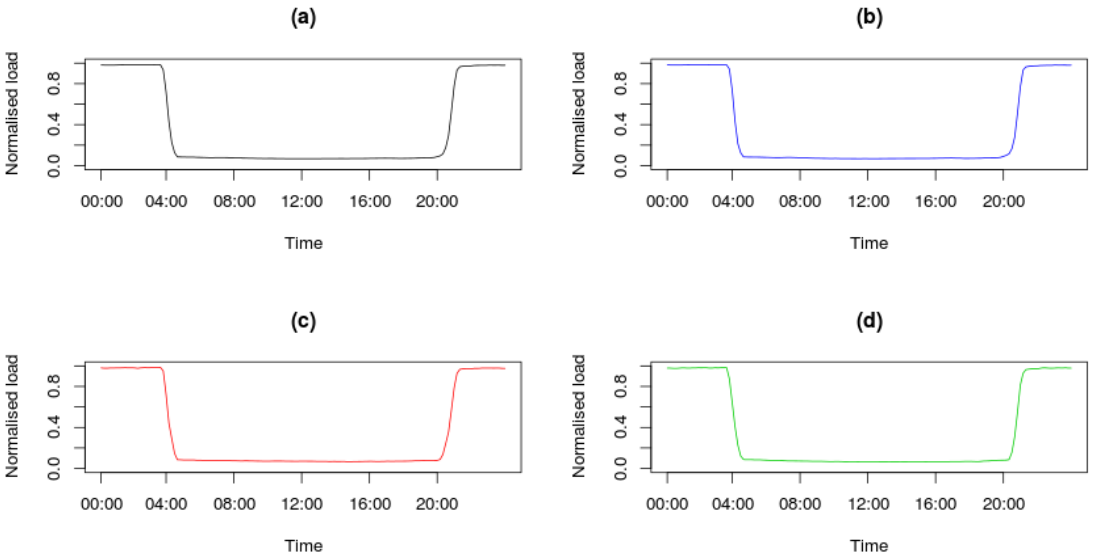

Figure C-19: Substation demand profiles for cluster 10 (summer): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

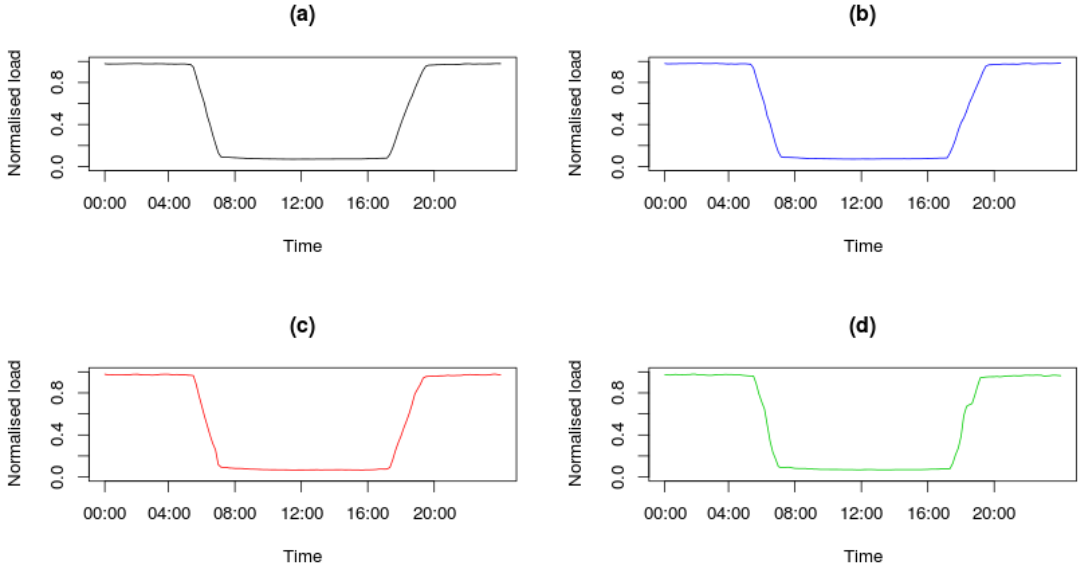

Figure C-20: Substation demand profiles for cluster 10 (autumn): Panels show results for (a) all days, (b) weekdays, (c) Saturdays and (d) Sundays.

Appendix. D

Outlier Analysis

The template errors can be caused by scaling factors and templates respectively. However, due to the significant variation in substation sizes, the same errors in templates are zoomed to different level. The error tends to be larger when the transformer rating increases. Another issue is that a hard threshold is required to clearly identify outliers. Outlier analysis aims to i) quantify the errors in terms of transformer rating of each substation; ii) to set threshold as maximum acceptable error, and to filter outliers based on the threshold.

The peak-error profile of a substation could be obtained by comparing the substation's real peak with the estimated value. And this fixed value has been used as peak-error quantitative index, which is then compared with the substation tolerance (maximum accepted error).To quantify the shape distortion, the maximum daily change of the shape-error profile has been assigned as the quantitative index.

The detailed analysis process has been illustrated in Figure D-1 below. For a substation n in cluster i, the obtained shape-error index and peak-error index are compared with a percentage of the substation's capacity, which is termed as maximum accepted error here. After going through this process for each substation in cluster i, the total numbers of substations that exceed peak-error tolerance and shape-error tolerance could be identified.

When designing LV networks, DNOs usually allow a probability level of 10% of exceeding the design load due to economic reasons. For templates assessment, it is reasonable to take 10% of transformer rating as threshold.

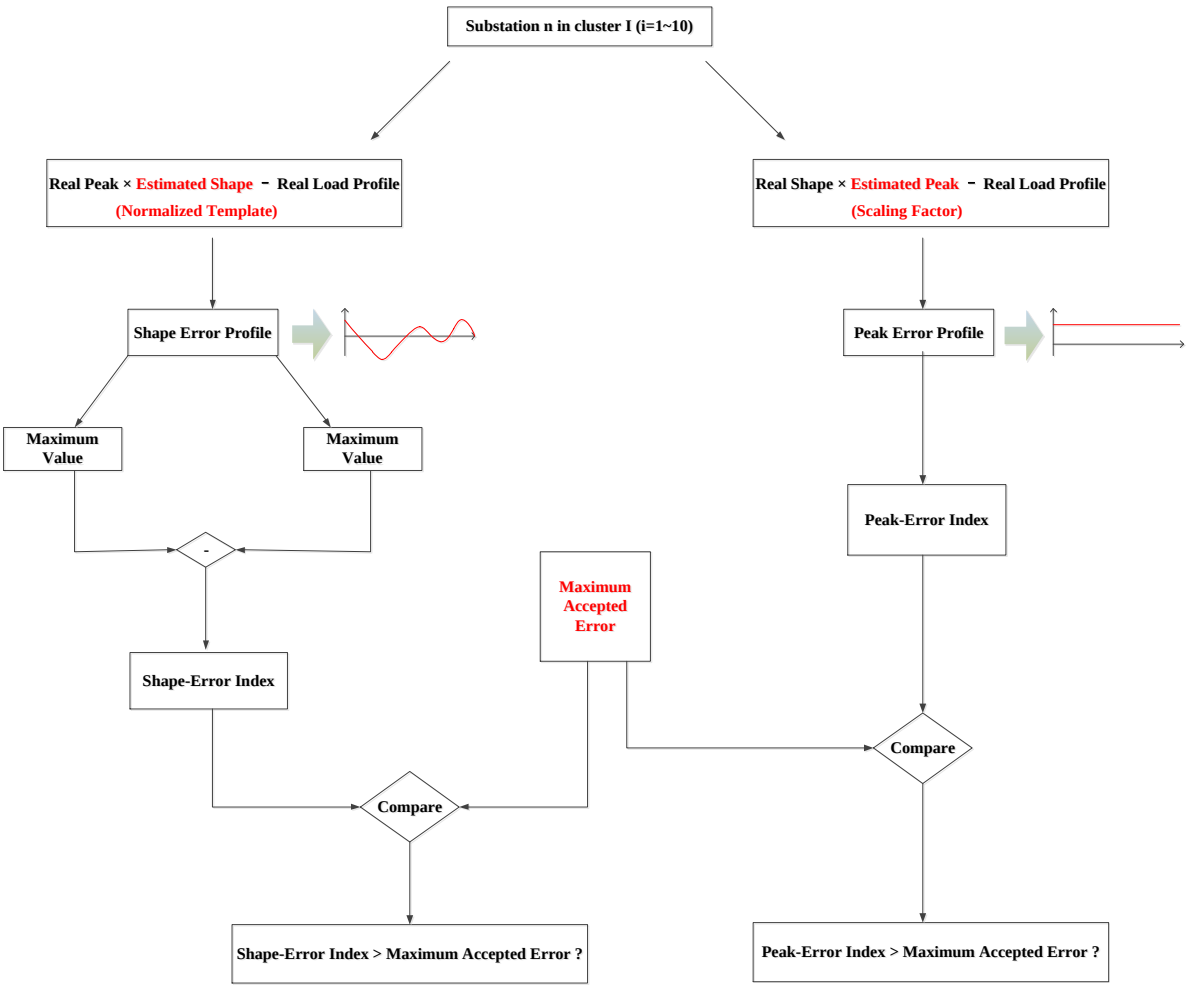

Figure. D-1. Assessment process

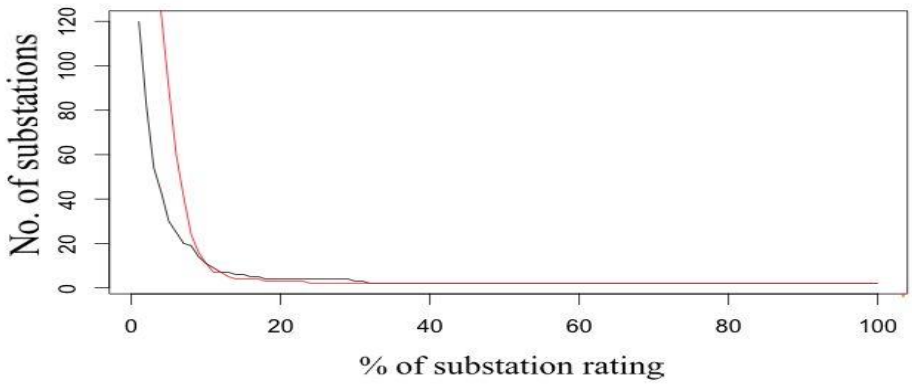

Figure. D-2. No. of substations exceeding different tolerance levels (cluster 4)

In Figure D-2, cluster 4 is taken as an example. The x axis means the maximum accepted error, which is a percentage of substation normal rating, and the y axis

shows the number of substations, whose error index exceeds the maximum accepted error. In this chart, the red line is shape-error situation, while the black one shows the peak-error situation.

Appendix. E

R-based Code of spectral decomposition

Note: due to confidential reason, the detailed code of LV network templates and MRC cannot be provided, but a simplified R-based code for decomposition and reconstruction assessment is provided in this part.

```
customer.average.fft=function(j=1000){

mean.customer=as.numeric(apply(subset(kresults,Station==j)[,3:50],2,mean))/max(as.numeric(apply(subset(kresults,Station==j)[,3:50],2,mean)))

indivi=as.numeric(subset(kresults,Station==j)[1,3:50])

mean.fft=fft(mean.customer)

indivi.fft=fft(indivi)

# extract magnitudes and phases

magn <- Mod(mean.fft)[1:25] # sqrt(Re(test)*Re(test)+Im(test)*Im(test))

phase <- Arg(mean.fft)[1:25] # atan(Im(test)/Re(test))

x=seq(0,47,by=1)

y=list()

y[[1]]=((magn[1]+magn[25])/48)*cos(0*x*pi/24+phase[1])

for (i in 2:24){

y[[i]]=(magn[i]/24)*cos((i-1)*x*pi/24+phase[i])

}

result.mean=matrix(NA,ncol=4,nrow=24)

for (m in 2:24){

est=y[[1]]

for (i in 2:m){

est=est+y[[i]]
```

```
}

result.mean[m,1]=abs(max(est)-max(mean.customer))/max(mean.customer)######peak magnitude error index (PMEI)

result.mean[m,2]=max(abs((est-mean.customer)))######Maximum Magnitude Error (MME)######

result.mean[m,3]=mean(abs((est-mean.customer)))######Mean absolute percentage Error (MME)######

result.mean[m,4]=min(abs(which(est==max(est))-which(mean.customer==max(mean.customer))))###peak time error(PTE) HALF-HOURLY###

}

magn2 <- Mod(indivi.fft)[1:25] # sqrt(Re(test)*Re(test)+Im(test)*Im(test))

phase2 <- Arg(indivi.fft)[1:25] # atan(Im(test)/Re(test))

x=seq(0,47,by=1)

y=list()

y[[1]]=((magn2[1]+magn2[25])/48)*cos(0*x*pi/24+phase2[1])

for (i in 2:24){

 y[[i]]=(magn2[i]/24)*cos((i-1)*x*pi/24+phase2[i])

}

result.indivi=matrix(NA,ncol=4,nrow=24)

for (m in 2:24){

 est=y[[1]]

 for (i in 2:m){

  est=est+y[[i]]

 }

 result.indivi[m,1]=abs(max(est)-max(indivi))/max(indivi)######peak magnitude error index (PMEI)

 result.indivi[m,2]=max(abs((est-indivi)))######Maximum Magnitude Error (MME)######

 result.indivi[m,3]=mean(abs((est-indivi)))######Mean absolute percentage Error (MAPE)######
```

```
 result.indivi[m,4]=min(abs(which(est==max(est))-which(indivi==max(indivi))))###peak time error(PTE) HALF-HOURLY###

}

output=list(magn,phase,magn2,phase2, result.mean, result.indivi)

return(output)

}

#rm(list= ls()[!(ls() %in% c('data','kresults','working.array.flat3b_nor_week'))])

##################wavelets  ################################################

library(waveslim)

library(Matrix)

library(chron)

#########################

customer.average.wave=function(j=1000){

mean.customer=as.numeric(apply(subset(kresults,Station==j)[,3:50],2,mean))/max(as.numeric(apply(subset(kresults,Station==j)[,3:50],2,mean)))

indivi=as.numeric(subset(kresults,Station==j)[1,3:50])

#######################################

x=mean.customer

y=indivi

  ma <- mra2(x, wf="haar",J=3, boundary="periodic",method="dwt")

 x.wt <- dwt(x, "haar", 3, "periodic")

 for (p in 1:4){

   x.wt[[p]][which(abs(x.wt[[p]])<0.1)]=0

 }

 est=apply(matrix(unlist(ma), nrow=48), 1, sum)

  ###################asess###

result.mean=matrix(NA,ncol=4,nrow=1)

 result.mean[,1]=abs(max(est)-max(mean.customer))/max(mean.customer)######peak magnitude error index (PMEI)
```

```
 result.mean[,2]=max(abs((est-mean.customer)))######Maximum Magnitude Error
(MME)######

 result.mean[,3]=mean(abs((est-mean.customer)))######Mean absolute percentage Error
(MME)######

 result.mean[,4]=min(abs(which(est==max(est))-
which(mean.customer==max(mean.customer))))###peak time error(PTE) HALF-HOURLY###

###### coefficients#######

 size=nnzero(x.wt$s3)+nnzero(x.wt$d1)+ nnzero(x.wt$d2)+nnzero(x.wt$d3)

##########################################################################
#######

ma <- mra2(y, wf="haar",J=3, boundary="periodic",method="dwt")

x.wt <- dwt(y, "haar", 3, "periodic")

for (p in 1:4){

 x.wt[[p]][which(abs(x.wt[[p]])<0.1)]=0

}

est=apply(matrix(unlist(ma), nrow=48), 1, sum)

####################asess###

result.ind=matrix(NA,ncol=4,nrow=1)

result.ind[,1]=abs(max(est)-max(y))/max(y)######peak magnitude error index (PMEI)

result.ind[,2]=max(abs((est-y)))######Maximum Magnitude Error (MME)######

result.ind[,3]=mean(abs((est-y)))######Mean absolute percentage Error (MME)######

result.ind[,4]=min(abs(which(est==max(est))-which(y==max(y))))###peak time error(PTE)
HALF-HOURLY###

###### coefficients#######

size2=nnzero(x.wt$s3)+nnzero(x.wt$d1)+ nnzero(x.wt$d2)+nnzero(x.wt$d3)

##########################################################################
##

  output=list(result.mean,size,result.ind,size2)

return(output)

 # x.s[i,52]=sum(x[i,3:50] - apply(matrix(unlist(ma), nrow=48), 1, sum))^2
```

```
}

##################aseess all wavelets reconstucted###############

assess.wave=matrix(NA,nrow=length(unique(kresults[,1])),ncol=11)

assess.wave[,1]=as.numeric(unique(kresults[,1]))

for (i in 1:length(unique(kresults[,1]))){

  output=customer.average.wave(j=unique(kresults[,1])[i])

  assess.wave[i,2:5]=output[[1]]

  assess.wave[i,6]=output[[2]]

  assess.wave[i,7:10]=output[[3]]

  assess.wave[i,11]=output[[4]]

  cat(i, "\t")

}
```

Publications

Journal Publications

R. Li; C. Gu; F. Li; G.Shaddick; “Development of Low Voltage Network Template Part I- Load Profile Clustering and Substation Classification”, *IEEE Transaction on Power Systems*, (accepted and ready to appear, No. : TPWRS-00313-2014).

R. Li; C. Gu; F. Li; G.Shaddick; “Development of Low Voltage Network Template Part II- Peak Load Estimation by Clusterwise Regression”, *IEEE Transaction on Power Systems*, (accepted and ready to appear, No. : TPWRS-00314-2014).

R. Li; Z. Wang; Chenghong Gu; Furong Li; H. Wu; "Time of Use Tariff Design for Domestic Customers by Model-based Clustering," *Applied Energy Special Issue 2014 ICAE*, (under review).

R. Li; F. Li; N.Smith; “Big Data Analysis for Smart Metering on Spectral Domain--Part I: Assessment of Spectral Analysis Techniques”, *IEEE Transaction on Smart Grid*, (under review).

R. Li; F. Li; N. Smith; “Big Data Analysis for Smart Metering on Spectral Domain--Part II: Multi-resolution Load Profile Clustering”, *IEEE Transaction on Smart Grid*, (under review).

J.Li; R. Li; C. Gu; R. Bhakar; F.Li; “A Time-of-Use Transmission Use of System Charging Methodology” *IEEE Transaction on Power Systems*, (under review).

Conference Publications

R. Li; Z. Wang; S. Blond and F. Li; " Development of Time-of-Use Price by Clustering Techniques," *Power and Energy Society General Meeting, 2014 IEEE* , July 2014

R. Li; G. Shaddick; H. Yan and F. Li; " Sample Size Determination of Photovoltaic by Assessing Regional Variability," *CIRED Workshop, Challenges of Implementing Active Distribution System Management*, June 2014

D. Shi; R. Li; R. Shi and F. Li; " Analysis of the Relationship between Load Profile and Weather Condition," *Power and Energy Society General Meeting, 2014 IEEE* , July 2014

Z. Wang; R. Li; Chenghong Gu; Furong Li; "Time of Use Tariff Design for Domestic Customers from Flat Rate by Model-based Clustering," *6th International Conference on Applied Energy*, 2014

S. Blond; R. Li; Z. Wang; F. Li; " Cost and emission savings from the deployment of variable electricity tariffs and advanced domestic energy hub storage management," *Power and Energy Society General Meeting, 2014 IEEE* , July 2014

R. Li; C. Gu; Y. Zhang; F. Li, "Implementation of load profile test for electricity distribution networks," *Power and Energy Society General Meeting, 2012 IEEE* , vol., no., pp.1,6, 22-26 July 2012

Reference

[1] "2050 Pathways Analysis," DECC, LondonJuly,2010 July,2010.

[2] S. Haben, M. Rowe, D. V. Greetham, P. Grindrod, W. Holderbaum, B. Potter, *et al.*, "Mathematical solutions for electricity networks in a low carbon future," in *Electricity Distribution (CIRED 2013), 22nd International Conference and Exhibition on*, 2013, pp. 1-4.

[3] N. M. Pearsall, K. M. Hynes, C. Martin, and M. Munzinger, "Analysis of Performance Parameters for UK Domestic PV Systems," in *Photovoltaic Energy Conversion, Conference Record of the 2006 IEEE 4th World Conference on*, 2006, pp. 2300-2303.

[4] "UK Renewable Energy Roadmap," DECC, LondonJuly 2011.

[5] L. Ran, G. Chenghong, Z. Yan, and L. Furong, "Implementation of load profile test for electricity distribution networks," in *Power and Energy Society General Meeting, 2012 IEEE*, 2012, pp. 1-6.

[6] E. Technology, "Assessing the Impact of Low Carbon Technologies on Great Britain's Power Distribution Networks," Energy Networks Associations2012.

[7] Y. Zhang, F. Li, Z. Hu, and G. Shaddick, "Quantification of Low Voltage Network Reinforcement Costs: A Statistical Approach," *Power Systems, IEEE Transactions on*, vol. 28, pp. 810-818, 2013.

[8] W. Zhimin, G. Chenghong, L. Furong, P. Bale, and S. Hongbin, "Active Demand Response Using Shared Energy Storage for Household Energy Management," *Smart Grid, IEEE Transactions on*, vol. 4, pp. 1888-1897, 2013.

[9] "Demand Side Response: Conflict between Eupply and Network Driven Optimisation," DECC2010.

[10] "Further consultation on implementing the Discretionary Funding Mechanism under the Low Carbon Networks Fund," Ofgem21 Aus 2014.

[11] "CDCM model user manual," energynetworks association2013.

[12] "Electricity User Load Profiles by profile class," Elexon. Ltd,1997.

[13] X. C, "Physical nature and exact definition of electric power," presented at the CPEM '90 Digest., Conference, 1990.

[14] Load Profiles and their use in Electricity Settlement [Online]. Available: http://www.elexon.co.uk/wp-content/uploads/2013/11/load_profiles_v2.0_cgi.pdf

[15] "Report on the Computer Program DEBUTE for the Design of LV Radial Networks," 1988.

[16] "Smart meters: a guide," Department of Energy & Climate Change22 January 2013.

[17] *Transition to smart meters*. Available: https://www.ofgem.gov.uk/electricity/retail-market/market-review-and-reform/smarter-markets-programme

[18] "Managing big data for smart grids and smart meters," USAMay 2012.

[19] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers," *Power Systems, IEEE Transactions on*, vol. 22, pp. 1120-1128, 2007.

[20] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *Power Systems, IEEE Transactions on*, vol. 21, pp. 933-940, 2006.

[21] Z. Tiefeng, Z. Guangquan, L. Jie, F. Xiaopu, and Y. Wanchun, "A New Index and Classification Approach for Load Pattern Analysis of Large Electricity Customers," *Power Systems, IEEE Transactions on*, vol. 27, pp. 153-160, 2012.

[22] Z. Shiyin and K. Tam, "A Frequency Domain Approach to Characterize and Analyze Load Profiles," *Power Systems, IEEE Transactions on*, vol. 27, pp. 857-865, 2012.

[23] S. Valero, M. Ortiz, C. Senabre, C. Alvarez, F. J. G. Franco, and A. Gabaldon, "Methods for customer and demand response policies selection in new electricity markets," *Generation, Transmission & Distribution, IET*, vol. 1, pp. 104-110, 2007.

[24] H. L. Willis, A. E. Schauer, J. E. D. Northcote-Green, and T. D. Vismor, "Forecasting Distribution system Loads Using Curve Shape Clustering," *Power Apparatus and Systems, IEEE Transactions on*, vol. PAS-102, pp. 893-901, 1983.

[25] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *Power Systems, IEEE Transactions on*, vol. 18, pp. 381-387, 2003.

[26] M. Espinoza, C. Joye, R. Belmans, and B. DeMoor, "Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic Time Series," *Power Systems, IEEE Transactions on*, vol. 20, pp. 1622-1630, 2005.

[27] J. Nazarko, A. Jurczuk, and W. Zalewski, "ARIMA models in load modelling with clustering approach," in *Power Tech, 2005 IEEE Russia*, 2005, pp. 1-6.

[28] S. V. Allera, N. D. Alcock, and A. A. Cook, "Load research in a privatised electricity supply industry," in *Metering Apparatus and Tariffs for Electricity Supply, 1990., Sixth International Conference on*, 1990, pp. 1-5.

[29] V. Hamidi, "Domestic Demand Response to Increase the Value of Wind Power," The Department of Electronic and Electrical Engineering, University of Bath, 2009.

[30] I. Drezga and S. Rahman, "Input variable selection for ANN-based short-term load forecasting," *Power Systems, IEEE Transactions on*, vol. 13, pp. 1238-1244, 1998.

[31] V. Ojala, "An integrated PC-program for the tariff planning of electric utilities and for national price statistics on electricity," in *Metering Apparatus and Tariffs for Electricity Supply, 1992., Seventh International Conference on*, 1992, pp. 72-76.

[32] C. S. Chen, J. C. Hwang, and C. W. Huang, "Application of load survey systems to proper tariff design," *Power Systems, IEEE Transactions on*, vol. 12, pp. 1746-1751, 1997.

[33] "Metering, load profiles and settlement in deregulated markets," System Tariff Issues Working Group, EURELECTRIC 2000-220-0004, 2000.

[34] C. L. Brooks, "A Stochastic Preference Technique for Allocation of Distribution Loads," presented at the Proc. American Power Conference, Univ. of Illinois, 1978.

[35] H. Willis and J. Aanstoos, "Some unique signal processing applications in power system planning," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, pp. 685-697, 1979.

[36] H. L. Willis, A. E. Schauer, J. E. D. Northcote-Green, and T. D. Vismor, "Forecasting Distribution System Loads Using Curve Shape Clustering," *Power Engineering Review, IEEE*, vol. PER-3, pp. 31-31, 1983.

[37] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," *Power Systems, IEEE Transactions on*, vol. 16, pp. 798-805, 2001.

[38] H. Shyh-Jier and S. Kuang-Rong, "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations," *Power Systems, IEEE Transactions on*, vol. 18, pp. 673-679, 2003.

[39] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: a review and evaluation," *Power Systems, IEEE Transactions on*, vol. 16, pp. 44-55, 2001.

[40] A. Khotanzad, R. Afkhami-Rohani, and D. Maratukulam, "ANNSTLF-Artificial Neural Network Short-Term Load Forecaster generation three," *Power Systems, IEEE Transactions on*, vol. 13, pp. 1413-1422, 1998.

[41] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer Classification and Load Profiling Method for Distribution Systems," *Power Delivery, IEEE Transactions on*, vol. 26, pp. 1755-1763, 2011.

[42] J. A. Jardini, H. P. Schmidt, C. M. V. Tahan, C. C. B. de Oliveira, and S. U. Ahn, "Distribution transformer loss of life evaluation: a novel approach based on daily load profiles," *Power Delivery, IEEE Transactions on*, vol. 15, pp. 361-366, 2000.

[43] S. V. a. H. Allera, A.G, "Load profiling for engergy trading and settlements in the UK electricity markets," in *DistribuTECH Europe DA/DSM*, London, UK, 1998.

[44] "Electricity user load profiles by profile class," UK ERC Energy Data Center Browse Archive; Electricity AssociationNov-97.

[45] T. Jonassen, "Opening of the Power Market to End Users in Norway 1991 - 1999," 1998.

[46] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *Generation, Transmission and Distribution, IEE Proceedings-*, vol. 151, pp. 395-400, 2004.

[47] J. A. Jardini, C. M. V. Tahan, M. R. Gouvea, S. U. Ahn, and F. M. Figueiredo, "Daily load profiles for residential, commercial and industrial low voltage consumers," *Power Delivery, IEEE Transactions on*, vol. 15, pp. 375-380, 2000.

[48] C. S. Chen, J. C. Hwang, Y. M. Tzeng, C. W. Huang, and M. Y. Cho, "Determination of customer load characteristics by load survey system at Taipower," *Power Delivery, IEEE Transactions on*, vol. 11, pp. 1430-1436, 1996.

[49] C. S. Chen, M. S. Kang, J. C. Hwang, and C. W. Huang, "Implementation of the load survey system in Taipower," in *Transmission and Distribution Conference, 1999 IEEE*, 1999, pp. 300-304 vol.1.

[50] E. Bompard, Carpaneto, E., Chicco, G., Napoli, R., Piglione, F., Postolache, P., and Scutariu, M, "Stratified sampling of the electricity customers for setting up a load profile survey," in *Proc. EuroConf. on Risk Management in Power Systems Planning and Operation Market Environment (RIMAPS)*, Funchal, 25–27 September 2000.

[51] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Electric energy customer characterisation for developing dedicated market strategies," in *Power Tech Proceedings, 2001 IEEE Porto*, 2001, p. 6 pp. vol.1.

[52] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "An approach to customers daily load profile determination," in *Power Engineering Society Summer Meeting, 2002 IEEE*, 2002, pp. 587-591 vol.1.

[53] B. D. Pitt and D. S. Kitschen, "Application of data mining techniques to load profiling," in *Power Industry Computer Applications, 1999. PICA '99. Proceedings of the 21st 1999 IEEE International Conference*, 1999, pp. 131-136.

[54] R. F. Chang and C. N. Lu, "Load profile assignment of low voltage customers for power retail market applications," *Generation, Transmission and Distribution, IEE Proceedings-*, vol. 150, pp. 263-267, 2003.

[55] "Report on the Design of Low Voltage Underground Networks for New Housing," 1986.

[56] J. Nazarko, R. P. Broadwater, and N. I. Tawalbeh, "Identification of statistical properties of diversity and conversion factors from load research data," in *Electrotechnical Conference, 1998. MELECON 98., 9th Mediterranean*, 1998, pp. 217-220 vol.1.

[57] C. f. E. R. (CER). CER Smart Metering Project [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[58] J. N. Fidalgo, M. A. Matos, and M. T. Ponce de Leão, "Assessing error bars in distribution load curve estimation," in *Artificial Neural Networks — ICANN'97*. vol. 1327, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, Eds., ed: Springer Berlin Heidelberg, 1997, pp. 1017-1022.

[59] "South Somerset population estimates for 2002," S. C. Council, Ed., ed, December 2009.

[60] "Household projections to 2031, England," ed. Department for communities and local government, 2009.

[61] "Regional and local authority electricity consumption statistics," L. Department of Energy & Climate Change, Ed., ed, 2008.

[62] "Report on the Draft Common Distribution Charging Methodology," Energy Networks Association, London August, 2009.

[63] A. M. M. Kociołek, M. Strzelecki P. Szczypiński "Discrete wavelet transform – derived features for digital image texture analysis," in *Proc. of Interational Conference on Signals and Electronic Systems*, Lodz, Poland, 2001, pp. 163-168.

[64] L. M. P. M. G. Pollitt, "The Economics of Energy (and Electricity)," Electricity Policy Research GroupApril 2011.

[65] M. G. Pollitt, Bialek, Janusz, "Electricity network investment and regulation for a low carbon future," Faculty of Economics, University of Cambridge, UKoct-2007.

[66] *Low Carbon Networks Project*. Available: http://www.westernpower.co.uk/About-our-Network/Low-Carbon-Networks-Project.aspx

[67] D. Infield and F. Li, "Integrating micro-generation into distribution systems — a review of recent research," in *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, 2008, pp. 1-4.

[68] S. V. Allera, and Horsburgh, A.G.:, "Load profiling for energy trading and settlements in the UK electricity markets," in *Proc. Conf. DistribuTECH Europe DA/DSM*, London, UK, 27–29 October 1998.

[69] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Load pattern-based classification of electricity customers," *Power Systems, IEEE Transactions on,* vol. 19, pp. 1232-1239, 2004.

[70] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *Power Systems, IEEE Transactions on,* vol. 20, pp. 596-602, 2005.

[71] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks," *Power Systems, IEEE Transactions on,* vol. 20, pp. 548-555, 2005.

[72] S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. G. Franco, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps," *Power Systems, IEEE Transactions on,* vol. 21, pp. 1672-1682, 2006.

[73] (2012). *LV Network Templates*. Available: http://www.ofgem.gov.uk/Networks/ElecDist/lcnf/stlcnp/year1/lv-network-templates/Pages/index.aspx

[74] M. Dale, "LV Network Templates for a Low-carbon Future," 2014.

[75] F. Murtagh, "A Survey of Recent Advances in Hierarchical Clustering Algorithms," *The Computer Journal,* vol. 26, pp. 354-359, November 1, 1983 1983.

[76] D. Müllner, "fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python," *Journal of Statistical Software,* vol. 53(9), pp. 1-18, 2013.

[77] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif., 1967, pp. 281-297.

[78] D. W. Hosmer and S. Lemeshow, *Applied logistic regression* vol. 354: Wiley-Interscience, 2000.

[79] A. D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," *Power Systems, IEEE Transactions on,* vol. 5, pp. 1535-1547, 1990.

[80] A. Sargent, R. P. Broadwater, J. C. Thompson, and J. Nazarko, "Estimation of diversity and kWHR-to-peak-kW factors from load research data," *Power Systems, IEEE Transactions on,* vol. 9, pp. 1450-1456, 1994.

[81] J. Nazarko and W. Zalewski, "The fuzzy regression approach to peak load estimation in power distribution systems," *Power Systems, IEEE Transactions on,* vol. 14, pp. 809-814, 1999.

[82] T. Konjic, V. Miranda, and I. Kapetanovic, "Fuzzy inference systems applied to LV substation load estimation," *Power Systems, IEEE Transactions on,* vol. 20, pp. 742-749, 2005.

[83] H. Ying-Yi and C. Zuei-Tien, "Development of energy loss formula for distribution systems using FCN algorithm and cluster-wise fuzzy regression," *Power Delivery, IEEE Transactions on,* vol. 17, pp. 794-799, 2002.

[84] B. Zhang, "Regression clustering," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 451-458.

[85] J. W. Taylor, "Short-Term Load Forecasting With Exponentially Weighted Methods," *Power Systems, IEEE Transactions on*, vol. 27, pp. 458-464, 2012.

[86] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*: Society for Industrial and Applied Mathematics, 1974.

[87] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 569-575, 2010.

[88] "Estimated impacts of energy and climate change policies on energy prices and bills," D. o. E. C. Change, Ed., ed, 2010.

[89] K. Imamura, N. Kubo, and H. Hashimoto, "Automatic moving object extraction using x-means clustering," in *Picture Coding Symposium (PCS), 2010*, 2010, pp. 246-249.

[90] A. J. R. Reis and A. P. Alves da Silva, "Feature extraction via multiresolution analysis for short-term load forecasting," *Power Systems, IEEE Transactions on*, vol. 20, pp. 189-198, 2005.

[91] A. S. Pandey, D. Singh, and S. K. Sinha, "Intelligent Hybrid Wavelet Models for Short-Term Load Forecasting," *Power Systems, IEEE Transactions on*, vol. 25, pp. 1266-1273, 2010.

[92] C. Ying, P. B. Luh, G. Che, Z. Yige, L. D. Michel, M. A. Coolbeth, *et al.*, "Short-Term Load Forecasting: Similar Day-Based Wavelet Neural Networks," *Power Systems, IEEE Transactions on*, vol. 25, pp. 322-330, 2010.

[93] Z. A. Bashir and M. E. El-Hawary, "Applying Wavelets to Short-Term Load Forecasting Using PSO-Based Neural Networks," *Power Systems, IEEE Transactions on*, vol. 24, pp. 20-27, 2009.

[94] K. Jungsuk, J. Flora, and R. Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data," *Smart Grid, IEEE Transactions on*, vol. 5, pp. 420-430, 2014.

[95] M. P. Tcheou, L. Lovisolo, M. V. Ribeiro, E. A. B. da Silva, M. A. M. Rodrigues, J. M. T. Romano, *et al.*, "The Compression of Electric Signal Waveforms for Smart Grids: State of the Art and Future Trends," *Smart Grid, IEEE Transactions on*, vol. 5, pp. 291-302, 2014.

[96] L. Sankar, S. R. Rajagopalan, S. Mohajer, and H. V. Poor, "Smart Meter Privacy: A Theoretical Framework," *Smart Grid, IEEE Transactions on*, vol. 4, pp. 837-846, 2013.

[97] D. Engel, "Wavelet-based load profile representation for smart meter privacy," in *Innovative Smart Grid Technologies (ISGT), 2013 IEEE PES*, 2013, pp. 1-6.

[98] A. M. Davood Rafiei "Efficient Retrieval of Similar Time Sequences Using DFT," in *Proc. Int'l Conf. Foundations of Data Organizations and Algorithms*.

[99] G. McLachlan and D. Peel, *Finite mixture models*: Wiley. com, 2004.

[100] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression," *Journal of classification*, vol. 5, pp. 249-282, 1988.

[101] P. Chaussé, "Computing Generalized Method of Moments and Generalized Empirical Likelihood with R," *Journal of Statistical Software*, vol. 34, May 2010.

[102] W. Z. Asad Hasan, Alireza S. Mahani. Fast Estimation of Multinomial Logit Models: R Package mnlogit [Online]. Available: http://arxiv.org/abs/1404.3177
[103] W. S. Nocedal J, *Numerical Optmization*, 2 ed.: Springer-Verlag, 2000.
[104] T. R. Hastie T, Friedman J, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.*, 2 ed.: Springer-Verlag, 2009.